Volume 84, Number 1, February 1976

Michael R. Darby: Three-and-a-Half Million U.S. Employees Have Been Mislaid: Or, an Explanation of Unemployment, 1934–1941

David F. Burgess: Tariffs and Income Distribution: Some Empirical Evidence for the United States

Anders Klevmarken and John M. Quigley: Age, Experience, Earnings, and Investments in Human Capital

J. C. Murphy and S. K. Das: Money Illusion and Balance-of-Payments Adjustment

Timothy A. Deyak and V. Kerry Smith: The Economic Value of Statute Reform: The Case of Liberalized Abortion

Harry G. Johnson: Destabilizing Speculation: A General Equilibrium Approach

Vincent J. Tarascio: Pareto: A View of the Present through the Past

The University of Chicago Press

JOURNAL OF
POLITICAL ECONOMY

Edited by
JACOB A. FRENKEL SAM PELTZMAN
HARRY G. JOHNSON GEORGE J. STIGLER
In cooperation with OTHER MEMBERS of the DEPARTMENT OF
ECONOMICS and the GRADUATE SCHOOL OF BUSINESS
of the UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES
(SEE INSIDE BACK COVER)

Editorial Assistants DAWN G. KAHN and EMILY SCHMIDT

The Journal of Political Economy is published bimonthly in February, April, June, August, October, and December by The University of Chicago Press. Subscription rates, U.S.A.: institutions, 1 year $20.00, 2 years $36.00, 3 years $55.00; individuals, 1 year $15.00, 2 years $27.00, 3 years $38.25. Student subscription rate, U.S.A.: 1 year $12.00 (letter from professor must accompany subscription). Other countries add $1.50 for each year's subscription to cover postage. Single copy rates: institutions $4.00, individuals $3.20. Supplements: institutions $4.50, individuals $3.50. Back issues are available from 1962 (vol. 70). Make all remittances payable to *Journal of Political Economy*, The University of Chicago Press, in United States currency or its equivalent.

Claims for missing numbers should be made within the month following the regular month of publication. The publishers expect to supply missing numbers free only when losses have been sustained in transit and when the reserve stock will permit.

Business correspondence should be addressed to The University of Chicago Press, 5801 Ellis Avenue, Chicago, Illinois 60637.

Letters to the editors and manuscripts should be addressed to the Editor of the *Journal of Political Economy*, 1126 East 59th Street, Chicago, Illinois 60637. **Manuscripts should be submitted in duplicate (3 copies if possible), accompanied by a $25.00 submission fee made payable to the Journal.** The fee will be refunded to authors whose papers are accepted for publication in the *Journal*. The proceeds from the submission fees are used to pay for refereeing services. Accepted manuscripts must be typed according to the University of Chicago *Manual of Style*. References should be typed double-spaced at the end of the article. Footnotes should be numbered in sequence and double-spaced following the references. Tables should follow the footnotes. Originals of the figures, drawn in india ink, should be submitted if the manuscript is accepted. Abstracts not exceeding 100 words should be submitted in duplicate along with the manuscript.

Applications for permission to quote from this *Journal* should be addressed to The University of Chicago Press and will be freely granted. Reproduction of articles requires written permission from the publisher and the author.

Reprinted volumes 1-72 available from Walter J. Johnson, Inc., 355 Chestnut Street, Norwood, New Jersey 07648. Volumes available **in microfilm** from University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106, **in microfiche** from Johnson Associates, P.O. Box 1017, Greenwich, Connecticut 06830.

Notice to subscribers: If you change your address, please notify us and your local postmaster immediately, giving *both* your old and your new address. *Allow four weeks for the change.*

Second-class postage paid at Chicago, Illinois, and at additional mailing office
© 1976 by The University of Chicago. All rights reserved
ISSN 0022-3808
PRINTED IN U.S.A.

Journal of Political Economy

Volume 84 Number 1 February 1976

Articles

1 Three-and-a-Half Million U S Employees Have Been Mislaid Or, an Explanation of Unemployment, 1934–1941
Michael R Darby

17 Tariffs and Income Distribution Some Empirical Evidence for the United States
David F Burgess

47 Age, Experience, Earnings, and Investments in Human Capital
Anders Klevmarken and John M Quigley

73 Money Illusion and Balance-of-Payments Adjustment
J C Murphy and S K Das

83 The Economic Value of Statute Reform The Case of Liberalized Abortion
Timothy A Deyak and V Kerry Smith

101 Destabilizing Speculation A General Equilibrium Approach
Harry G Johnson

109 Pareto A View of the Present through the Past
Vincent J Tarascio

Review Article

123 Economics in Dialectical Dialect
Martin Bronfenbrenner

Communications

131 Some Evidence on the Factor Content of West Germany's Foreign Trade
Robert M Stern

143 The Property Rights Theory of the Firm Empirical Results from a Natural Experiment
H E Frech III

153 Thornton's Estimates of Efficiency Losses in Soviet Industry Some Fixed-Point-Method Recalculations
John Whalley

161 Sharing Risks of Deferred Payment
Steven Shavell

169 An Inquiry into the Economics of Race-Track Gambling
Arthur Gruen

179 Currency Denominations and the Price Level
Chau-nan Chen

Book Reviews

185 George W Douglas and James C Miller III, *Economic Regulation of Domestic Air Transport Theory and Policy*
William A Jordan

189 Ray C Fair, *A Model of Macroeconomic Activity* Vol 1 *The Theoretical Model*
Uri M Possen

192 Richard B McKenzie and Robert J Staaf, *An Economic Theory of Learning*
Edward Lazear

194 John H A Munro, *Wool, Cloth, and Gold The Struggle for Bullion in Anglo-Burgundian Trade, 1340–1478*
Frank Spooner

195 Luigi L Pasinetti, *Growth and Income Distribution Essays in Economic Theory*
Duncan K Foley

198 Lars G Sandberg, *Lancashire in Decline A Study in Entrepreneurship, Technology, and International Trade*
Donald N McCloskey

200 A Michael Spence, *Market Signaling Informational Transfer in Hiring and Related Screening Processes*
Emmett Keeler

201 Jeffrey G Williamson, *Late Nineteenth Century American Development A General Equilibrium History*
Gavin Wright

389 **A Note on Tariffs, Nontariff Barriers, and Labor Protection in United States Manufacturing Industries**
John H Cheh
395 **Import Instability and Tariffs: Some Welfare Implications of Price Stabilization**
Toshihisa Toyoda
401 The Leff Conjecture: Some Contrary Evidence
John R Hanson II
407 A Note on Risk, Return, and Equilibrium Empirical Tests
Menachem Brenner

Miscellany

411 Toward a Deeper Economics of Sleeping
T C Bergstrom

Book Reviews

413 Michael Allingham, *Equilibrium and Disequilibrium. A Quantitative Analysis of Economic Interaction*
R M Witcomb
415 Michael Parkin and Michael T Sumner, Eds , *Incomes Policy and Inflation*
Edgar L Feige
417 Almarin Phillips, Ed , *Promoting Competition in Regulated Markets*
Michael M Murphy
420 Martin F J Prachowny, *Small Open Economies Their Structure and Policy Environment*
Harry G Johnson
420 E D Steele, *Irish Land and British Politics Tenant-Right and Nationality, 1865–1870*
Barbara L Solow
422 Jan Tinbergen, *Income Distribution: Analysis and Policies*
Barry R Chiswick

Effective July 1, 1976, all manuscripts submitted to the *Journal of Political Economy* must be accompanied by a submission fee of $30 00 (instead of the present $25 00) The submission fees will continue to be used to meet the fees to referees, which will also be raised on July 1

THE EDITORS

Journal of Political Economy

Volume 84 Number 2 April 1976

Articles

207 A Classical Macroeconometric Model for the United States
Thomas J. Sargent

239 Economic Development in Communist China
Colin Clark

265 A Stochastic Model of Applied Research
Robert E. Evenson and Yoav Kislev

283 The Marginal Cost of Public Funds
Edgar K. Browning

299 A Model of Optimal Human-Capital Accumulation and the Wages of Young High School Graduates
Charles Brown

317 Determining Legislative Preferences on the Minimum Wage An Economic Approach
Jonathan I. Silberman and Garey C. Durden

Communications

331 Perceived Wealth in Bonds and Social Security. A Comment
Martin Feldstein

337 Barro on the Ricardian Equivalence Theorem
James M. Buchanan

343 Reply to Feldstein and Buchanan
Robert J. Barro

351 Earnings and Individual Variations in Postschool Human Investment
Charles B. Knapp and W. Lee Hansen

359 The Conundrum of the Low Issue of National Bank Notes
John A. James

369 Autocorrelated Growth Rates and the Pareto Law: A Further Analysis
Daniel R. Vining, Jr.

381 It Takes Two to Tango, or *Sind* "Separable Externalities" ***Überhaupt*** *Moglich?*
William J. Baumol

Journal of Political Economy

Volume 84 Number 3 June 1976

Articles

427 Inflation Uncertainty and Expected Returns on Treasury Bills
Eugene F Fama

449 Human Capital and Labor Supply A Synthesis
Alan S Blinder and Yoram Weiss

473 An Economic Analysis of Eminent Domain
Patricia Munch

499 Economically Rational Expectations Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?
Edgar L Feige and Douglas K Pearce

523 Uncertainty, Waiting Time, and Capacity Utilization A Stochastic Theory of Product Quality
Arthur De Vany

543 Project Evaluation, Shadow Prices, and Trade Policy
Ronald Findlay and Stanislaw Wellisz

553 National Health Insurance Some Costs and Effects of Mandated Employee Coverage
Bridger M Mitchell and Charles E Phelps

Review Article

573 The New Contractarians
Scott Gordon

Communications

591 The Effect of Unemployment Compensation on a Seasonal Industry Agriculture
Barry R. Chiswick

603 A Dynamic Model of the Firm
Arvind Virmani

615 **Multilateral and Bilateral Measures of Effective Exchange Rates in a World Model of Traded Goods**
Stanley W Black

623 Triangular Graphs
Joe Latham

631 The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics
Thomas J Sargent

641 Measuring Gains and Losses
George W McKenzie

647 The Effects of Property Taxes and Local Public Spending on Property Values A Theoretical Comment
Bruce W Hamilton

Book Reviews

651 Walter C Labys, ed , *Quantitative Models of Commodity Markets*
Charles C Cox

652 M Ishaq Nadiri and Sherwin Rosen, *A Disequilibrium Model of Demand for Factors of Production*
Frank Brechling

Effective July 1, 1976, all manuscripts submitted to the *Journal of Political Economy* must be accompanied by a submission fee of $30 00 (instead of the present $25 00) The submission fees will continue to be used to meet the fees to referees, which will also be raised on July 1

THE EDITORS

Journal of Political Economy

Volume 84 Number 4, Part 1 August 1976

Articles

655 Economies of Scale in U.S. Electric Power Generation
Laurits R. Christensen and William H. Greene

677 A Strategic Analysis of Monetary Interdependence
Koichi Hamada

701 Migration, Age, and Education
Aba Schwartz

721 Price Dispersion in the Government Securities Market
Kenneth D. Garbade and William L. Silber

741 Some Quantitative Aspects of the Economics Journal Literature
Richard E. Quandt

757 The Origin and Development of Media of Exchange
Robert A. Jones

777 Alternative Stabilization Policies for Less-developed Economies
Basant K. Kapur

797 Direct versus Indirect Remedies for Externalities
Jerry Green and Eytan Sheshinski

Review Article

809 Technology, Growth, and History
Jeffrey G. Williamson

Communications

821 The Friedman-Savage Utility Function in Cross-cultural Perspective
Frederic L. Pryor

835 Fertility Interactions and Modernization Turning Points
Paul R. Gregory and J. M. Campbell, Jr.

849 Testing the Monetary Approach to Devaluation in Developing Countries
Michael Connolly and Dean Taylor

861 Another Advantage of the Division of Labor
Nathan Rosenberg

869 Taxes, Spending, and Property Values. A Comment and Further Results
George Richard Meadows

881 Do Import Levies Matter? The Case of Sweden
Gary Sampson and Alexander Yeats

893 Money, Interest, and Prices in Market Disequilibrium. A Comment
George H Blackford

Book Reviews

895 Karl F Helleiner, *Free Trade and Frustration. Anglo-Austrian Negotiations, 1860–1870*
Victoria Curzon

897 Odin Knudsen and Andrew Parnes, *Trade Instability and Economic Development An Empirical Study*
R M Sundrum

899 Thomas Sowell, *Classical Economics Reconsidered*
Samuel Hollander

Journal of Political Economy

Volume 84 Number 5 October 1976

Articles

903 The Cobb-Douglas Production Function Once Again. Its History, Its Testing, and Some New Empirical Values
Paul H Douglas

917 A Statistical Theory of Expenditures in Capital Maintenance and Repair
George C Bitros

937 Temporary Layoffs in the Theory of Unemployment
Martin Feldstein

959 Interfuel Substitution in Steam Electric Power Generation
Scott E Atkinson and Robert Halvorsen

979 Money and Economic Activity in the Open Economy. The United Kingdom, 1880–1970
P D Jonson

1013 Private Pensions and Saving· New Evidence
Alicia H Munnell

1033 The Economics of Information and Retail Gasoline Price Behavior: An Empirical Analysis
Howard P. Marvel

1061 A Theory of Government Enterprise
Cotton M Lindsay

1079 Exhaustible Resources and Industrial Structure. A Nash-Cournot Approach to the World Oil Market
Stephen W. Salant

Communications

1095 The Effect of Broadcasting on Political Campaign Spending: An Empirical Investigation
Burton A. Abrams and Russell F. Settle

1109 Advertising, Concentration, and Price-Cost Margins
Allyn D Strickland and Leonard W Weiss

1123 Market Organization and the Effect of Import Quotas Empirical Evidence
Susumu Hondai

1133 A Purchasing-Power-Parity Application to Demand for Money in Afghanistan
Maxwell J Fry

Book Reviews

1139 Henry G Manne and Roger Leroy Miller, eds , *Gold, Money and the Law*
Elmus R Wicker

1141 Jørgen Pedersen, *Essays in Monetary Theory and Related Subjects*
Jürg Niehans

1142 Arnold Plant, *Selected Economic Essays and Addresses*
Almarin Phillips

Journal of Political Economy

Volume 84 Number 6 December 1976

Dr. Zakir Husain Library
Jamia Nagar
New Delhi-110025

Articles

1145 A New and Superior Process for Making Social Choices
T Nicolaus Tideman and Gordon Tullock

1161 Expectations and Exchange Rate Dynamics
Rudiger Dornbusch

1177 An Alternative Approach to the Analysis of Taxation
Yoram Barzel

1199 The Successes and Failures of Professor Smith
George J Stigler

1215 Futures Trading and Market Information
Charles C Cox

1239 Consequences and Causes of Public Ownership of Urban Transit Facilities
B Peter Pashigian

1261 The Incidence and Efficiency Effects of Taxes on Income from Capital
John B Shoven

Communications

1285 The Incidence and Efficiency Effects of Taxes on Income from Capital A Reply
Arnold C Harberger and Neil Bruce

1293 Optimum Replacement of Capital Goods with Labor-saving Technical Progress A Comparison of the Early New England and British Textile Firm
Peter L Swan

1305 Some More Problems with Income-contingent Loans· The Case of Medical Education
Roger Feldman

1313 Policy Implication of Full-Loss-Offset Capital Gains Tax on Forward Contracts: A Note on Leland's "Optimal Forward Exchange Positions"
Dan Galai

1319 The Social Costs of Monopoly and Regulation: Some Comments
Göran Skogh

1325 Distribution, the Aggregate Consumption Function, and the Level of Economic Development. Some Cross-Country Results
Philip A. Della Valle and Noriyoshi Oguchi

1335 Firm Demand for Information about Price and Technology
Richard E Kihlstrom

1343 The Desirability of Cheating Incentives and Randomness in the Optimal Income Tax
Laurence Weiss

1353 On the Specification of the Demand for Money. The Real Rate of Return versus the Rate of Inflation
Benjamin Eden

1361 Vindication of a "Common Mistake" in Welfare Economics
John C Panzar and Robert D Willig

1365 Rejoinder
Robert Dorfman

Book Reviews

1367 Michael J Cullen, *The Statistical Movement in Early Victorian Britain The Foundations of Empirical Social Research*
Phyllis Deane

1368 Susan Howson, *Domestic Monetary Management in Britain, 1919–1938*
David Laidler

1370 Marvin H Kosters in association with J Dawson Ahalt, *Controls and Inflation. The Economic Stabilization Program in Retrospect*
John P Gould

1372 Tibor Scitovsky, *The Joyless Economy An Inquiry into Human Satisfaction and Human Dissatisfaction*
James W. Friedman

1374 Peter Temin, *Did Monetary Forces Cause the Great Depression?*
E S Shaw

1378 Adrian Wood, *A Theory of Profits*
Eugene Lerner

1381 Index to Volume 84

Three-and-a-Half Million U.S. Employees Have Been Mislaid: Or, an Explanation of Unemployment, 1934–1941

Michael R Darby

University of California, Los Angeles, and National Bureau of Economic Research

A major conceptual error in the standard BLS and Lebergott unemployment estimates for 1930–43 is reported Emergency workers (employees of government contracyclical programs such as WPA) were counted as unemployed on a normal-jobs-to-be-created instead of job-seekers unemployment definition For 1933–41, the corrected unemployment levels are reduced by 2–3 5 million people and the rates by 4–7 percentage points The corrected data show strong movement toward the natural unemployment rate after 1933 and are very well explained by an anticipations-search model using annual full-time earnings

When Keynesian or post-Keynesian economists denounce the central monetarist proposition that the economy converges to full employment in the absence of shocks, the prime empirical contradiction is the well-known persistence of high (though falling) levels of unemployment from 1934 through 1941 If convergence to full employment is so slow, it is argued, it might as well be ignored So the explanation of this era is the central challenge to monetarist or equilibrium economists Despite

---

Perceptive questioning by James R Lothian and comments by Anna Schwartz wrested the admission that my theoretical framework could not explain the high unemployment rates of 1934 through 1941 unless millions of people were gainfully employed as "unemployed " Thinking of generous unemployment insurance, I had no idea that my hyperbole would prove literally correct Mrs Schwartz's incomparable knowledge of data resources provided the key references that put me on the trail Valuable comments on an earlier draft were received from Robert Barro, Harold Demsetz, Milton Friedman, and Robert Lucas The author hopes that Bob Gordon will forgive the obvious recycling of his original (1969) title and that this recycling does not seriously reduce the total stock of titles (see Darby 1973) The regressions were estimated by Nurhan Helvacian on the NBER's Troll system This article was written while the author was Harry Scherman Research Fellow at the National Bureau of Economic Research but is not an official report of the National Bureau

[*Journal of Political Economy*, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

immense theoretical and empirical effort, no real answer has been given

This paper argues that the inability of monetarist models to explain such large, persistent unemployment rates is an asset because those unemployment rates did not in fact exist The unemployment rates reported for the period count as unemployed literally millions of employees on the payrolls of government emergency relief projects such as the Works Progress Administration and the Civilian Conservation Corps Both modern search and Keynesian theory require that contracyclical government employment not be counted as unemployment

Section I demonstrates that the standard unemployment series[1] do indeed count "emergency workers" as unemployed and suggests the reasoning which led to this misconception Corrected data for employment, unemployment, and the unemployment rate are presented in Section II It is shown in Section III that the corrected data are consistent with a search approach to unemployment in which deviations of the actual from the natural level of unemployment are explained by deviations of actual from anticipated levels of wages and prices

I. The Standard Treatment of Emergency Workers

There are two standard data sets for the labor force, employment, and unemployment from 1929 through 1943, the official Bureau of Labor Statistics (BLS) data and the Lebergott data taken from Stanley Lebergott's book (1964) Either series could be called the "Lebergott data," since the original presentation of the BLS data is in Lebergott (1948) [2] Though both standard data sets thus can be attributed to Lebergott, the treatment of "emergency workers" as unemployed was standard practice in the estimates of that period [3]

Lebergott (1948)—perhaps too inured to the Census Bureau definition —makes no mention whatsoever of the exclusion of emergency workers from employed and their inclusion in unemployed This practice can be confirmed only by tracing it back to the Census Bureau source document for the 1940–46 data to which the 1929–39 estimates are tied [4] I found only one explicit statement by Lebergott on the inclusion of emergency workers among the unemployed:

---

[1] Specific reference is to the series B1 (Lebergott) and B2 (Bureau of Labor Statistics) reported in U S Department of Commerce (1973)

[2] Lawrence R Klein was listed as Chief, Office of Publications of the BLS

[3] Most notably, the Bureau of the Census and the National Industrial Conference Board (NICB) A typical census publication reads "On the basis of the answers to these questions, persons in the labor force were classified into the following employment status categories (*a*) 'Employed (except on public emergency work),' including those reported as at work and those with a job but not at work, and (*b*) unemployed, including those 'On public emergency work' and those 'Seeking work' " (U S Bureau of the Census 1944, p 2) Similarly, "Unemployment—The difference between the number of persons in the labor force and the number of persons employed Persons on the payrolls of government emergency relief projects, such as WPA, CCC, and out-of-school NYA, are not considered employed" (NICB 1945, p 389)

> These estimates for the years prior to 1940 are intended to measure the number of persons who are totally unemployed, having no work at all For the 1930's this concept, however, does include one large group of persons who had both work and income from work—those on emergency work In the United States we are concerned with measuring lack of regular work and do not minimize the total by excluding persons with made work or emergency jobs This contrasts sharply, for example, with the German practice during the 1930's when persons in the labor force camps were classed as employed, and Soviet practice which includes employment in labor camps, if it includes it at all, as employment [5]

If we exclude the forced analogy of the WPA to Nazi labor camps and suppose that Lebergott did nothing more than follow common practice in the 1930s, the question remains whether there is any valid rationale for this treatment

The answer appears to depend upon the particular year in question From 1930 through 1932 work relief was used as an apparently minor part of the general relief program of some states and cities,[6] although this increased in the latter part of 1932 under Reconstruction Finance Corporation advances (repayment later waived) "for relief and work relief to needy and distressed people and in relieving the hardships resulting from unemployment "[7] Employment under these programs was of a limited, part-time nature, with average annual wages per worker of

---

[4] U S Bureau of the Census 1947 At p 1 it states "During the period in which public emergency work projects were being conducted by the Works Project Administration (WPA), the National Youth Administration (NYA), the Civilian Conservation Corps (CCC) and State and local work relief agencies (this period ended about June 1943), persons at work on, or assigned to, such projects were also included among the unemployed " I suspect that almost no one currently at the BLS is aware that they are reporting (for example, U S Department of Labor 1975, p. 19) data for 1933–43 which includes emergency workers among the unemployed Certainly the standard explanatory notes (pp 135–42 for the same issue) state that "*employed persons* comprise (*a*) all those who during the survey week did any work at all as paid employees, . " count people currently on similar government programs as employed, and mention no problems under "Historic Comparability "

[5] Lebergott (1957, p 229), repeated verbatim in Lebergott (1964, pp 184–85)

[6] The national income accounts show annual state and local work-relief wages (in millions of dollars) as 1930, 4, 1931, 46, 1932, 92, 1933, 341, 1934, 680, 1935, 791, 1936, 74, 1937, 48, 1938, 8, 1939, 10, 1940, 10, 1941, 6; 1942, 2 (U.S Department of Commerce 1947, table 14, p 27)

[7] Emergency Relief and Construction Act of July 21, 1932, Title I, Sec. 1a.

$155.87, or only 13 percent of the [illegible] com equivalent (FTE) employee in all [illegible] of F wages in agriculture. So a reasonable case could certainly be made f counting these "underemployeds" as unemployed even though it inconsistent with both current practice and the treatment of the wages the national income accounts

Substantial federal subsidies for state relief (including work relie continued through November of 1933 State work-relief employees ar estimated to have averaged 1 72 million people in 1933.[8] This implies a annual average wage based on the national income accounts data— $197 80 or about 19 percent of the average annual FTE earnings in al industries So although enrollment in state programs increased in thi year from about 1 percent to about 3 percent of the labor force,[9] it i still not unreasonable to count state and local work-relief employees a unemployed

Various New Deal programs created what is known as the emergenc government labor force Broadly speaking, these emergency workers car be divided into two classes 1 those working in youth (16 24 years old programs under the Civilian Conservation Corps (CCC, April 1933-Jun 1943) and the out-of-school National Youth Administration (NYA, January 1936 May 1943 programs (2) The much larger regular program consisting of the Civil Works Administration (CWA, November 1933 July 1934), the Emergency Work-Relief Program (April 1934-December 1935 of the Federal Emergency Relief Administration (FERA), and the Works Progress Administration (WPA, July 1935 June 1943 titled Work Projects Administration beginning July 1939) The regular programs, which employed on average 82 percent of the emergency labor force, were mainly 75 85 percent of expenditures) involved in construction projects The FERA's Emergency Work-Relief Program was operated through state agencies, but the others were directly administered by the federal government

The emergency government labor force worked on qualitatively different terms than the state work-relief employees Jobs were substantially full time (24 40 hours per week) and at competitive wages[10] Preference was given to those eligible for relief, but this was not a bar to unemployed people The programs were mainly aimed at and attractive to unskilled labor From 1934 through 1940 annual wages were about one-half of the annual average earnings per full-time employee in all

---

[8] U S Department of Commerce 1966, p 98

[9] The U S Department of Commerce estimates of the number of state and local work-relief workers for 1930-32 are 20,000, 299,000, and 592,000 (1966, p 98)

[10] The conditions of the major programs are summarized in appendices 2 5 of the U S National Resources Planning Board (1942) Wages were generally set at "prevailing" or "fair" levels for the skill class of the work, or at minimum rates where they were higher than the market wage

TABLE 1

EMERGENCY GOVERNMENT LABOR FORCE EMPLOYMENT AND WAGES, 1933–43

YEAR	FEDERAL EMERGENCY WORKERS (THOUSANDS) (1)	AVERAGE ANNUAL WAGES		RELATIVE WAGE OF EMERGENCY WORKERS (%) (4)
		Emergency Workers ($) (2)	All Industry ($) (3)	
33	471	755 83	1,048	72 1
34	2,475	508.69	1,091	46 6
35	2,575	469.51	1,137	41.3
936	3,653	595 13	1,184	50 3
937	2,707	605 10	1,258	48 1
938	3,572	590 71	1,230	48 0
939	3,216	583 02	1,264	46 1
940	2,792	565 54	1,300	43 5
941	2,192	553 38	1,443	38 3
942	909	638 06	1,709	37 3
943	85	588 24	1,951	30 2

SOURCES —Col 1 Federal work relief full-time and part-time employees, table 6 3, line 78, pp 98, 99, of U S Department of Commerce (1966) For 1934 and 1935, 1,335 thousand and 1,585 thousand, respectively, employees of FERA's state-administered Emergency Work-Relief Program are added on the basis of annual average of monthly data for April 1934–December 1935 in table 13, 3d col , p 46, of U S Work Projects Administration (1942) Slightly different data (taken from preliminary reports such as FERA [1936] but showing monthly breakdowns by program) on the emergency government labor force are available in NICB (1941, 1945) Col 2 Wages and salaries of federal work relief employees, table 6 2, line 78, pp 94, 95, of U S Department of Commerce (1966) divided by col 1 For 1934 and 1935 $495 million and $598 million, respectively, have been added to wages and salaries as the proportionate share of FERA workers in total state and local work relief wages reported (ibid , line 84) for those years This is a minimum estimate as *work relief recipients in state and local programs in 1933 averaged less than half the average for 1934* Col 3 Average annual earnings per full-time employee in all industries, Series D696, U S Bureau of the Census (1960), p 95 Col 4 Col 2/col 3

industries, as shown in table 1 So although the classification of state and local work-relief employees is debatable, there is little room for doubt that the employees of the federal contracyclical programs should indeed be counted as employed

From the Keynesian viewpoint, labor voluntarily employed on contracyclical construction and other government projects should certainly be counted as employed On the search approach to unemployment, a person who accepts a job and withdraws from the activity of search is clearly employed The contracyclical intent of the government employer is irrelevant to the definition of unemployment as currently used by the BLS Whether unemployment would have been higher in the absence of contracyclical employment programs—or lower in the absence of restrictive monetary policy—is an entirely separate question from how many people without jobs are seeking one.

Besides bureaucratic inertia which carried forward the earlier treatment of state work-relief employees, the misclassification of emergency workers can be attributed to an implicit definition of unemployment as the difference between the normal labor force and those employed in normal jobs The pioneering estimates by the National Industrial Conference Board (NICB) aimed at estimating how many jobs would have to be created in order to get back to a normal situation of no depression and therefore no emergency government labor force Measuring jobs to be created leads to different treatment than measuring people without work who are seeking it [11] Of course, no consideration was given to emergency employment's crowding out private employment or the state and local construction work which it replaced

In summary, the treatment of government work-relief employees as unemployed in the BLS and Lebergott data is inconsistent with both the modern BLS definition of unemployment and with the U S national income accounts This treatment may be defensible in the case of state and local workers but not for the federal emergency government labor force [12]

II. Corrected Estimates for 1929–43

Since the misclassification of emergency workers as unemployed does not affect the civilian labor force but only its components, it is straightforward to correct the BLS and Lebergott data The corrected data presented in table 2 are consistent with the current definition which counts as employed "all those who during the survey week did any work at all as paid employees "[13] Comparing the standard[14] and corrected unemployment figures in columns 12 and 13, it is clear that the misclassification of emergency workers almost completely masked both the speed and strength of the return of unemployment toward normal levels in 1934 through 1941 The employment level for 1936 is understated in the standard data by over 3 5 million people and by over 2 million as late as 1941

This difference is seen most clearly in terms of the unemployment rates reported in table 3 For example, the Lebergott data overstate the unemployment rate by 1 2 percentage points in 1932 This rises to 4 3 in 1933, 5 8 in 1934, 5 9 in 1935, and a full 7 0 percentage points in 1936

---

[11] On this view employment can exceed the "normal" labor force, and negative unemployment is reported as in 1906, 1917–19, and 1942–44 (NICB 1945, pp 38–39) Milton Friedman suggested it might be interesting to compute the unemployment rate for the postwar era on the definition of unemployment as job seekers plus employees of "make-work" jobs Unfortunately it is much easier to estimate which make-work employees were counted as unemployed in the 1930s than to estimate the number of postwar government (and private?) employees on make-work jobs

[12] A word of warning should be given about the Census Bureau data on the duration of unemployment These, too, should be taken with a large grain of salt because "[t]he duration of unemployment represents the length of time (up to March 30, 1940) during which these persons had been seeking work or working on public emergency projects" (U S Bureau of the Census 1943, p 7)

[13] U S Department of Labor 1975, p 138

[14] Lebergott explicitly adopted (1964, pp 407–8) the BLS figures on unemployment for 1929–39 because his revised estimates differed only slightly from the BLS data already in widespread use To maintain consistency, Lebergott reported employment as his revised labor-force estimates less the BLS unemployment data From 1940 onward, Lebergott adopted the BLS data without change

TABLE 2

CIVILIAN LABOR FORCE, CLASSIFIED BY EMPLOYMENT STATUS
STANDARD AND CORRECTED DATA, 1929–43
(THOUSANDS)

YEAR	CIV LABOR FORCE		EMERGENCY WORKERS	EMPLOYMENT				UNEMPLOYMENT	
	Lebergott (5)	BLS (6)	(7)	Lebergott (8)	BLS (9)	Cor Leb (10)	Cor BLS (11)	BLS/Leb (12)	Corrected (13)
1929	47,757	49,180	0	46,207	47,630	46,207	47,630	1,550	1,550
1930	48,523	49,820	20	44,183	45,480	44,203	45,500	4,340	4,320
1931	49,325	50,420	299	41,305	42,400	41,604	42,699	8,020	7,721
1932	50,098	51,000	592	38,038	38,940	38,630	39,532	12,060	11,468
1933	50,882	51,590	2,195	38,052	38,760	40,247	40,995	12,830	10,635
1934	51,650	52,230	2,974	40,310	40,890	43,284	43,864	11,340	8,366
1935	52,283	52,870	3,087	41,673	42,260	44,760	45,347	10,610	7,523
1936	53,019	53,440	3,744	43,989	44,410	47,733	48,154	9,030	5,286
1937	53,768	54,000	2,763	46,068	46,300	48,831	49,063	7,700	4,937
1938	54,532	54,610	3,591	44,142	44,220	47,733	47,811	10,390	6,799
1939	55,218	55,230	3,255	45,738	45,750	48,993	49,005	9,480	6,225
1940	55,640	55,640	2,830	47,520	47,520	50,350	50,350	8,120	5,290
1941	55,910	55,910	2,209	50,350	50,350	52,559	52,559	5,560	3,351
1942	56,410	56,410	914	53,750	53,750	54,664	54,664	2,660	1,746
1943	55,540	55,540	85	54,470	54,470	54,555	54,555	1,070	985

SOURCES—Cols. 5, 8, and 12 Lebergott 1964, p 12, cols 6, 9, and 12 Lebergott 1948, p 51, col 7 federal, state, and local work relief employees, lines 78 + 84, table 6.3, pp 98–99, U.S Department of Commerce, 1966, col 10 col 8 + col 7, col 11 col 9 + col 7, col 13 col 5 – col 10 = col 6 – col 11 = col 12 – col 7

TABLE 3

UNEMPLOYMENT RATES
STANDARD AND CORRECTED DATA, 1929–43
(%)

	UNEMPLOYMENT RATES			
YEAR	Lebergott (14)	BLS (15)	Corrected Lebergott (16)	Corrected BLS (17)
1929	3.2	3.2	3.2	3.2
1930	8.9	8.7	8.9	8.7
1931	16.3	15.9	15.7	15.3
1932	24.1	23.6	22.9	22.5
1933	25.2	24.9	20.9	20.6
1934	22.0	21.7	16.2	16.0
1935	20.3	20.1	14.4	14.2
1936	17.0	16.9	10.0	9.9
1937	14.3	14.3	9.2	9.1
1938	19.1	19.0	12.5	12.5
1939	17.2	17.2	11.3	11.3
1940	14.6	14.6	9.5	9.5
1941	9.9	9.9	6.0	6.0
1942	4.7	4.7	3.1	3.1
1943	1.9	1.9	1.8	1.8

SOURCES.—Col. 14 = col. 12/col. 5, col. 15 = col. 12/col. 6, col. 16 = col. 13/col. 5, col. 17 = col. 13/col. 6.

NOTE.—If state and local work relief employees were instead classified as unemployed and only federal emergency workers (col. 1, table 1 above) were classified as employed, the corrected unemployment rates would be somewhat higher for the years 1931–40 although the only substantial difference is for 1933. These alternative data are:

Year	1931	1932	1933	1934	1935	1936	1937	1938	1939	1940
Corrected Lebergott	16.3	24.1	24.3	17.2	15.4	10.1	9.3	12.5	11.3	9.6
Corrected BLS	15.9	23.6	24.0	17.0	15.2	10.1	9.2	12.5	11.3	9.6

Not until 1941 was the overstatement down to even 3.9 percentage points. Put another way, the reported reduction of unemployment from 1932 to 1936 was only about half of the actual reduction.

The note to table 3 presents the unemployment rates corresponding to counting federal work-relief employees as employed but state and local work-relief employees as unemployed. The only substantial difference is for 1933.

The civilian labor force is estimated in both the standard and corrected data by a linear trend between decennial benchmarks adjusted for demographic shifts. Coen (1973) has argued that the civilian labor force varies procyclically because of the discouraged-worker phenomenon. Alternative unemployment rates—lower than those reported in table 3—would be estimated by applying Coen's or a similar procedure to the corrected employment data. Substantial questions arise however in applying relations estimated from postwar data to the interwar years, especially where data on wages and hours are involved (see below).

III. An Anticipations-Search Explanation of the Corrected Data

The corrected unemployment rates for 1934 through 1940 are still high, though no longer absurdly so. Since 1900 the unemployment rate shows "no discernible trend"[15] Except for the 1900s and 1930s, every decade in the twentieth century has shown an average unemployment rate of 5 percent[16]—about the current estimate of the natural rate of unemployment The question naturally arises whether the corrected data are consistent with a model in which the unemployment rate tends to the natural rate in the absence of shocks.

Robert Lucas and Leonard Rapping (1972) have shown that from 1930 through 1933 their simple equilibrium model (1969) does well both qualitatively and quantitatively[17] It does slightly better with the corrected data for 1933 Presumably, a similar result would hold for any anticipations-search model which can be expressed as

$$u_t = a_0 + a_1 \log (W_t/W_t^*) + a_2 \log (P_t/P_t^*), \qquad (1)$$

where u_t is the unemployment rate, a_0 is the natural unemployment rate, W_t and W_t^* are the actual and anticipated nominal wage levels, and P_t and P_t^* are the actual and anticipated price levels Lucas and Rapping report that their model fails to explain high levels of unemployment from 1934 through 1941 because rapid increases in nominal wages imply that the unemployment rate should fall to near the natural rate in 1934 or at the very latest by 1937 Even the corrected data show a much less rapid fall

The main problem derives from the hourly wage series used by Lucas and Rapping,[18] which rises by 12 percent in 1934 This reflects a 4 percent rise in annual compensation per full-time employee and a 7 2 percent fall in hours of work Other series on hourly wages show a similar jump in 1934 due to a large decline in reported hours and a small increase in annual wages[19] This is extremely peculiar behavior for the average hours of work, which is normally a very reliable leading indicator—leading more at peaks than at troughs Indeed Gerhard Bry's study (1959) of the average work week showed only one cycle in average weekly hours not corresponding to a reference cycle the peak of July 1933 and trough of September 1934

---

[15] Lebergott 1964, p 188

[16] Lebergott (ibid , p 189) reports a decade average of 4 percent for the 1900s and 18 percent for the 1930s The corrected data (cols 16 and 17 of table 3) show a 1930s average of 14 percent but 4 percent for the 1940s

[17] Lucas and Rapping used the Lebergott estimates of unemployment Their discussion starts in 1930 instead of 1929, apparently because of data availability in series not considered here They were fortunate in this regard as no reasonable model could explain the reported 3 2 percent unemployment rate in 1929—a drop of 1 percentage point from 1928 The 1930 unemployment rate of 8 9 percent seems reliable because of the decennial census, but it is incredible that a bit under 10 percent drop in real GNP could be associated with a 5 7 percentage point increase in the unemployment rate from 1929 to 1930

[18] U S Department of Commerce (1966) estimates of compensation of employees in all industries (pp 90–93, line 1) divided by the product of (*a*) the number of FTE employees in all industries (pp 102–5, line 1) and (*b*) the average annual hours per employee (Denison 1962, p 37 [not p 85], col 3) Unfortunately Lucas and Rapping (1969, p 742, n 27) had misinterpreted their (secondary) source as reporting "annual hours worked per year by full-time employees" when it in fact reports "actual average annual hours per employee" as is seen in the original presentation by James Knowles (1960, p 26)

[19] See for example, series such as D626 and D627 on wages and hours in manufacturing in U S Bureau of the Census (1960, p 92)

The explanation of this exceptional behavior of hours and wages is easy to find the National Industrial Recovery Act (passed June 16, 1933) and succeeding legislation The NRA codes and later wage and hours laws aimed at limiting the hours of work and increasing hourly wages Significantly, Bry reports "The sample reporting manhours expanded greatly between 1932 and 1934, largely because of cooperation by the various code authorities of the NRA" (1959, p 107) Now this reported fall in average hours and rise in wages could reflect some combination of three factors (1) wholehearted compliance with the law and codes, (2) a change in firms' reporting practices—entirely legal—so as to eliminate inclusion of rest and meal breaks in reported hours, and (3) falsification of reported hours and wages as part of an evasion scheme A cynical economist (such as I) might suspect that the data primarily reflect items 2 and 3, but this is clearly a topic for detailed historical research [20] If the change in hours is largely spurious, average earnings per FTE employee would more accurately reflect the development of wages in the 1930s

In any case, traditional models of search have used "the wage" as a shorthand expression for total conditions of employment This concept would appear to be better represented by annual earnings of FTEs—or of all employees—in a period in which total hours are being rationed

Average annual FTE earnings are compared with the Lucas and Rapping series on the money wage rate in table 4 Columns 20 and 21 are in index form for the benefit of eyeball econometricians The discrete jump in the Lucas and Rapping series for 1934 is not reflected in the average annual FTE earnings series although the two series otherwise generally conform Because of the upward shift in the Lucas and Rapping series, nominal wages are reported as slightly above their 1929 level in 1937 This does not occur until 1941 in the average annual FTE earnings series

In view of these questions concerning the use of reported hourly wages, equation (1) was estimated for 1930–41 using the corrected Lebergott

---

[20] Such as in the thesis of Michael Weinstein, currently under way at M I T It has been argued, most recently by Robert Hall (1975), that the NRA codes and the succeeding wage and hour laws raised wages by fiat and so changed the character of unemployment from search unemployment (1929–33) to minimum-wage unemployment (1934–41) This is a competing but possibly complementary hypothesis to that in the text Recent papers by Finis Welch (1974) and Jacob Mincer (1974) note that the main effect of such laws would be on the distribution of employment between covered and uncovered sectors rather than the unemployment rate per se Further research on the precise effects of these institutional changes is clearly called for

TABLE 4

COMPARISON OF AVERAGE ANNUAL EARNINGS PER FULL-TIME EMPLOYEE IN ALL INDUSTRIES WITH LUCAS AND RAPPING'S MONEY WAGE RATE

YEAR	Avg Annual FTE Earnings ($ per Year) (18)	Lucas and Rapping Wage Series ($ per Hour) (19)	Avg Annual FTE Earnings (1929 = 100) (20)	Lucas and Rapping Wage Series (1929 = 100) (21)
1929	1,405	56	100	100
1930	1,368	56	97	100
1931	1,275	53	91	95
1932	1,120	48	80	86
1933	1,048	46	75	82
1934	1,091	51	78	91
1935	1,137	52	81	93
1936	1,184	53	84	95
1937	1,258	57	90	102
1938	1,230	58	88	104
1939	1,264	58	90	104
1940	1,300	60	93	107
1941	1,443	66	103	118

SOURCES.—Col. 18: Average annual earnings per full-time employee in all industries, series D696, U.S. Bureau of the Census (1960), p. 95; col. 19: Lucas and Rapping (1972), p. 189; col. 20: col. 18/14.05; col. 21: col. 19/0.0056.

unemployment rate, average annual FTE earnings, and the GNP price deflator for u_t, W_t, and P_t, respectively.[21] The anticipated levels of wages and prices were estimated by a simple adaptive expectations formulation:

$$W_t^* = \lambda W_t + (1 - \lambda)(1 + g_1)W_{t-1}^*, \qquad (2)$$

$$P_t^* = \lambda P_t + (1 - \lambda)(1 + g_2)P_{t-1}^*. \qquad (3)$$

Here λ is the coefficient of adjustment and g_1 and g_2 are the trend growth rates of wages and prices.[22] More elaborate models of anticipations might well yield better explanations but were not tried because of the scarcity of degrees of freedom I use the word "anticipations" instead of "expectations" by way of reminder that short-run aggregate supply or Phillips curve reflects past expectations of current prices, as embodied in contracts, as well as current expectations of searchers (see Darby 1976, esp chap 14)

---

[21] The u_t data are from table 3, col. 16; the W_t series for 1929–41 are from table 4, col. 18. It was extended back to 1924 by using the Kuznets estimates (series D685, Bureau of the Census 1960, p. 95) and a ratio splice based on the 1929 overlap. This earnings series is $19–$21 less than average FTE compensation for 1930–35, but this difference rises to $50 by 1937 and $63 by 1941 because of the inclusion of employer social security and unemployment insurance taxes. This increase in employer taxes would not represent a wage increase from the point of view of a job-seeking worker, so the earnings series was used. The P_t series for 1929–41 is Commerce Department data (series B62, Department of Commerce 1973, pp. 222–23). It was extended back to 1924 by using the NBER-Kendrick estimates (series B61, ibid.) and a 1929 ratio splice.

[22] The required initial values of W_t^* and P_t^* were taken as equal to the actual values in 1924, a year in which Lebergott estimates a 5.0 percent unemployment rate. The anticipated growth rate of prices g_2 was taken as 0, which is consonant with both the gold standard and actual experience from 1921 to 1941. The corresponding anticipated growth rate of nominal (and real) wages was estimated as 1½ percent per annum. If these growth rates are underestimates, the constant term will be biased upward.

Equation (1) was estimated by a simple maximum-likelihood technique Separate OLS regressions were run conditional upon the value of λ for $\lambda = 0, 0\,025, 0\,05, \quad, 0\,95, 0\,975$ The maximum-likelihood estimate is regression 1 in table 5 The simple model does rather well at explaining the fluctuations in the unemployment rate over the period 1930–41 One suspects that the standard error of estimate and 91 percent R^2 must be overoptimistic The constant term is well below the sample mean of the dependent variable (13 1 percent) but is significantly above the natural rate of unemployment of about 5 percent The λ-estimate is lower than the Lucas and Rapping estimate which was inferred, however, from the coefficient of the lagged dependent variable in a Koyck transformation of equation (1) The independent variables are rather collinear, and it was observed that higher values of λ (within the confidence interval) increase the relative coefficient and significance of log (W_t/W_t^*) and, conversely, for lower values of λ So the reported standard errors and t-statistics for the constant and coefficients, being conditional on $\hat{\lambda}$, are misleading guides to significance However the F-test for R^2 rejects the null hypothesis that all the coefficients are zero at any reasonable level of significance Regressions 2 and 3 were estimated suppressing in turn the wage and price terms They are suggestive of a somewhat more rapid adjustment of anticipated wages than anticipated prices Regressions based on the unemployment rates in the note to table 3 (state and local work-relief employees counted as unemployed) differ only trivially from those reported in table 5

Equation (1) was also estimated over the entire period 1930-65 of the Lucas and Rapping (1969) study [23] Over this whole period the adaptive expectations approach is much more questionable, but so would be any expectations model The estimate is regression 4 in table 5 and is essentially identical with regression 3 for the depression period with the price term constrained to zero [24]

These results are best interpreted as negating the view that anticipations-search models have no role in explaining unemployment during 1934–41 Only a much broader study integrating postwar data, New Deal institutional changes, and more sophisticated expectations models could

---

[23] The unemployment rate and GNP deflator for 1944–65 are series B2 and B62 from U S Department of Commerce (1973) The average annual FTE earnings in all industries is line 1, table 6 5, pp 107–9 (ibid , 1966)

[24] A check of the residuals revealed three greater than 1 standard error of estimate during the depression period The unemployment rate is overpredicted by 3 0 percentage points in 1930 and by 1 7 in 1936, it is underpredicted by 4 0 percentage points in 1932

TABLE 5

Estimates of the Anticipation-Search Equation (1)

Regression	Period	Constant	Coefficient of log (W_t/W_t^*)	 log (P_t/P_t^*)	$\hat{\lambda}$	SEE	R^2 (cor)	D-W
1	1930–41	8 64 (0 73) 11 77	−7 61 (15 93) −0 48	−63 84 (21 14) −3 02	0 15 [0 07, 0 36]	1 48	9149	2 04
2	1930–41	8 83 (0 59) 15 10		−73 36 (6 73) −10 89	0 15	1 42	9219	2.11
3	1930–41	10 19 (0 58) 17 50	−65 51 (6 92) −9 47		0 30	1 62	8987	1.57
4	1930–65	10 04 (0 30) 33 36	−68 53 (9 61) −7 13	19 42 (12 22) 1 59	0 275	1 45	9231	1.12

NOTE.—The standard errors are given in parentheses below the coefficients and above the t-statistics The reported statistics have been adjusted for the loss of 1 df in fitting λ. A greater than 90% confidence interval on $\hat{\lambda}$ (based on the asymptotic distribution of the logarithm of the likelihood function) is given in brackets below the estimated value 0.15 for number 1 SEE = Standard error of estimate provide a clear picture as to the amount of excess unemployment which can be attributed to normal search behavior versus the amount caused by New Deal programs

IV. Conclusions and Summary

The corrected data radically alter our empirical conception of unemployment during the 1930s The fact that unemployment rose to nearly one quarter of the labor force in 1932–33 is unchanged, but that fact is readily explained by the heroic ineptness of Federal Reserve policy The extremely slow recovery from 1934 through 1941 is shown to be a fiction based on erroneous data From 1933 to 1936 the corrected unemployment rate fell by 3 6 percentage points per year, and there is every reason to suppose that the natural rate of about 5 percent would have been reached by 1938 had the Fed not doubled reserve requirements between August 1936 and May 1937

To be sure, during 1934–40 between 5 and 7 percent of the labor force was employed in contracyclical, make-work public construction jobs, but there is little empirical evidence in the postwar period that the government spending multiplier is much above zero over such a long period as 2 or 3 years Certainly the unemployment rate behaves as if the emergency programs crowded out regular employment—particularly on the state and local construction works which would otherwise have been required

The utter inability reported by Lucas and Rapping of anticipations-search models[25] to explain unemployment during the recovery period 1934–41 can be attributed in part to errors in the unemployment data and in part to the use of an inappropriate hourly wage plagued by the introduction of reporting biases with the advent of the NRA Only a broader study over an extended time period can provide a realistic estimate of the residual increase in unemployment that might be directly attributable to such programs as the NRA But is is clear that the cyclical behavior of unemployment in the 1930s generally conforms to movements in actual versus expected wages and prices as predicted by modern natural-rate models

In summary, a major conceptual error in the standard BLS and Lebergott data has been uncovered Emergency workers (employees of government contracyclical programs such as WPA) were counted as unemployed due to the use of an implicit definition of "normal-jobs-to-be-created" instead of the standard job-seekers definition Estimates were affected for 1930–43, and during 1933–41, by huge amounts—2–3 5 million unemployed people or about 4–7 percentage points on the unemployment rate. The corrected data reveal a strong movement toward the natural unemployment rate after 1933 A simple anticipations-search model of the Lucas-Rapping type explains the corrected data very well when average full-time earnings are used as the nominal wage, but further research is required on the possible influence of NRA and similar programs in increasing the unemployment rate

---

[25] Strictly speaking, their model should be termed an "equilibrium" or "cleared-labor-market" model, although it is in the standard anticipations-search form

References

Bry, G *The Average Workweek as an Economic Indicator* National Bureau of Economic Research Occasional Paper, no 69 New York Nat Bur Econ Res, 1959

Coen, R M "Labor Force and Unemployment in the 1920's and 1930's A Re-Examination Based on Postwar Experience" *Rev Econ and Statis* 55 (February 1973) 46–55

Darby, M R "Paper Recycling and the Stock of Trees" *J.P E* 81, no 5 (September/October 1973) 1253–55

——— *Macroeconomics The Theory of Income, Employment, and the Price Level* New York McGraw-Hill, 1976

Denison, E F *The Sources of Economic Growth in the United States and the Alternatives before Us* Committee for Economic Development Supplementary Paper, no 13 New York Committee Econ Development, 1962

Gordon, R J "$45 Billion of U S Private Investment Has Been Mislaid" *A E R* 59 (June 1969) 221–38

Hall, R E "Are Wages Determined by Supply and Demand?" Mimeographed Massachusetts Inst Tech, January 1975

Knowles, J W "The Potential Economic Growth of the United States" U S Congress, Joint Economic Committee Study Paper, no 20 Washington Government Printing Office, 1960

Lebergott, S "Labor Force, Employment, and Unemployment, 1929–1939 Estimating Methods" *Monthly Labor Rev* 66 (July 1948) 50–53

——— "Annual Estimates of Unemployment in the United States, 1900–1954" In *The Measurement and Behavior of Unemployment*, sponsored by Universities–National Bureau Committee for Economic Research National Bureau of Economic Research Special Conference Series, no 8 Princeton, N J Princeton Univ Press, 1957

——— *Manpower in Economic Growth The American Record since* 1800 New York· McGraw-Hill, 1964

Lucas, R E, Jr, and Rapping, L A "Real Wages, Employment, and Inflation" *J P E* 77, no 5 (September/October 1969) 721–54

——— "Unemployment in the Great Depression Is There a Full Explanation?" *J P E* 80, no 1 (January/February 1972). 186–91.

Mincer, J "Unemployment Effects of Minimum Wages" National Bureau of Economic Research Working Paper, no 39 New York. Nat. Bur Econ. Res, 1974

National Industrial Conference Board. *The Economic Almanac for 1941–42.* New York Nat Indus Conference Board, 1941.

——— *The Economic Almanac for 1945–46.* New York. Nat. Indus. Conference Board, 1945

U.S Bureau of the Census *16th Census of the United States Population The Labor Force (Sample Statistics), Industrial Characteristics.* Washington· Government Printing Office, 1943.

——— *16th Census of the United States Population Estimates of the Labor Force, Employment and Unemployment in the United States, 1940 and 1930.* Washington: Government Printing Office, 1944

——— "Labor Force, Employment and Unemployment in the United States, 1940 to 1946" *Labor Force Bulletin, Current Population Reports,* Series P-50, no 2. Washington Government Printing Office, n d (circa 1947)

——— *Historical Statistics of the United States, Colonial Times to 1957* Washington. Government Printing Office, 1960

U S Department of Commerce *National Income and Product Statistics of the United States, 1929–46* Supplement to *Survey of Current Business,* July 1947 Washington: Government Printing Office, 1947

——— *National Income and Product Accounts, 1929–1965* Supplement to *Survey of Current Business,* August 1966 Washington Government Printing Office, 1966

——— *Long-Term Economic Growth, 1860–1970* Washington Government Printing Office, 1973

U S Department of Labor *Employment and Earnings* 21 (April 1975) 19, 135–42

U S Federal Emergency Relief Administration *Statistical Summary of Emergency Relief Activities, January 1933 through December 1935* Washington Government Printing Office, circa 1936

U S National Resources Planning Board *Security, Work, and Relief Policies* Washington Government Printing Office, 1942

U S Work Projects Administration *Final Statistical Report of the Federal Emergency Relief Administration* Washington Government Printing Office, 1942

Welch, F "Minimum Wage Legislation in the United States" *Econ Inquiry* 12 (September 1974) 285–318

Tariffs and Income Distribution: Some Empirical Evidence for the United States

---

David F. Burgess

University of Western Ontario

The duality between transformation functions and joint cost functions and the generalized linear-generalized Leontief form are used to provide a framework suitable for assessing the income distributional effects of tariffs Imports are viewed as intermediate inputs which are combined with capital and labor by the aggregate private domestic producing sector to produce two composite outputs for final demand The manner by which higher tariffs affect factor rewards depends critically upon whether the technology is separable or nonjoint Both of these popularly used special cases are shown to be of questionable empirical validity for the postwar U S economy Nevertheless, both models predict that higher tariffs raise real wages

I. Introduction

How does the U S tariff structure affect the distribution of income and the returns to its primary factors, capital and labor? Who gains and who loses if U S tariffs are reduced or eliminated? Despite the importance of the question of factoral protection and an early impetus given to it by Stolper and Samuelson (1941) and Travis (1964), the evidence that economists have been able to uncover to shed light on this question is rather indirect, surprisingly weak, and somewhat contradictory The bulk of recent empirical research on tariffs has been addressed to the closely related question of the sectoral protection afforded by a tariff structure Stimulated by the introduction and refinement of the concept of effective protection, economists have attempted to measure the extent to which a particular tariff structure shifts the allocation of resources among industries The more ultimate and prior question of factoral protection has been considered more or less as a peripheral matter, if at all

The author wishes to thank, without implicating, Erwin Diewert, Harry Johnson, and Alan Woodland

[*Journal of Political Economy*, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

The effective rate of protection purports to measure the extent to which the returns to primary factors in particular sectors of the economy are altered by a tariff structure. A ranking of sectors by their "effective rates" is then supposed to provide an indication of how resources are reallocated as a consequence of the tariff. As Travis (1968) points out, if all primary factors were sector specific, then sectoral and factoral protection would coincide. However, in a longer-run perspective, the basic primary factors, capital and labor, are mobile across sectors. Therefore, at best, the effective rate gives some indication of short-run income distributional effects prior to any resource movements.

General equilibrium theorists have observed that when there are numerous individual traded commodities produced under competitive conditions and constant returns to scale, and when there are a few scarce but mobile primary factors, the pattern of production and trade is indeterminate for an economy (see Melvin 1968, Travis 1972). Therefore, it is impossible to infer the direction of change in resource allocation that results from a change in the tariff structure as pursued by proponents of effective protection. However, it remains possible to predict the change in factor prices and income distribution, since factor prices are determined by commodity prices under competitive conditions. Thus, since tariffs alter the domestic prices of commodities, the impact of a tariff structure on factor prices is, theoretically at least, a perfectly tractable problem.

Value added in any sector is just a weighted sum of the returns to the primary factors, where the weights are the per unit direct primary-factor demands. If the tariff structure tends to raise wage rates relative to rental rates, then those processes which are labor intensive should exhibit high effective rates. This line of reasoning led Basevi (1966) and Balassa (1965) to regress their measures of effective rates on corresponding measures of direct labor content for each sector. Finding no significant positive correlation for the United States, they concluded that the U.S. tariff fails to protect labor. The measures of effective rates calculated by the above authors were based upon the assumption of zero substitutability between material and primary inputs. This deficiency, together with the recognition that commodity prices and factor prices are functionally related by the matrix of direct and indirect primary-factor input coefficients, led Travis (1968) to regress nominal tariff rates on direct and indirect labor content. He found a positive and significant association, and therefore concluded that U.S. tariffs indeed protect labor.

Travis's results are intuitively reasonable, since we normally think of the United States as labor scarce in terms of unadjusted man-hours and capital abundant when human capital is included. However, while U.S. tariffs may protect sectors which have relatively high man-hour requirements, many of these sectors tend to be standardized processes with low educational and skill requirements (see Baldwin 1971). The wage component of value added for these sectors may therefore on balance be relatively low Since the ultimate general equilibrium effect of restrictions on imports is to reduce exports, and since these latter processes tend to require relatively large amounts of skilled labor, the overall impact of tariffs on wages and the return to capital in the United States is by no means clear The question of factoral protection afforded by the U S tariff gains further importance in view of the increasing international mobility of capital If the competitive return to capital is reduced by higher tariffs, then this may partly explain the persistent net capital outflows experienced by the United States over the postwar period

The major objective of this paper is to reinvestigate the question of factoral protection afforded by the U S tariff using a completely different but more direct, and hopefully more appropriate, methodology We begin by observing that at the level of the overall economy a variety of primary factors, capital and labor services plus imported goods, are inputs into the productive process for the delivery of various types of output to final demand (see Burgess 1974*b*) We can summarize this technology by a transformation function $T(y, x) = 0$ giving an implicit association between a vector of final output y and a vector of primary-factor input x which is technically capable of producing y and no more This representation obscures the intricate intersectoral flows of materials and primary-factor services that occur within the economy The focus is on primary factors, so that imported materials are explicitly represented while domestically produced material inputs "wash out" in the aggregation process Observe that the technology pertains to final demand outputs rather than outputs from individual sectors, much of which is material input into other sectors Finally, all imported goods enter the technology as inputs despite their stage of processing, since what we are ultimately interested in is how final demand is satisfied and all imported goods are a potential source for meeting this demand either directly or indirectly

In this paper, I assume that the aggregate private producing sector employs two primary factors, capital and labor services, along with imported materials to produce two categories of output for final demand The composite outputs are durables and nondurables on the one hand and nongovernment services and structures on the other This decomposition is suggested by the significant differences in the movements of their corresponding price indexes It is also not unduly unrealistic to regard durables and nondurables as predominantly tradable goods, while services and structures are essentially nontraded I treat the bundle of imported goods as a single composite good whose domestic price is the world price modified by the tariff If primary factors are in fixed supply at any moment, and if the producing sector takes output and imported input prices as given, any tariff increase will tend to encourage substitution away from imports to alternative inputs for production for final demand This will have implications for the prices of the primary-factor services and the distribution of income depending upon the specific form of the technology It is a major objective of this paper to develop these implications and test their empirical content

Before proceeding with this latter task, the highly aggregative approach to the question of factoral protection perhaps warrants further comment It is, of course, completely unsuitable for making any statements about sectoral protection since we ignore the intricacies of the tariff structure and the economy's input-output matrix No distinction is therefore made between a change in the domestic price of an imported natural-resource product and an imported consumer durable, for example While proponents of effective protection might recoil at this, we emphasize that our interest is not sectoral but factoral protection With small exceptions, any type of import has a close, if not perfect, domestically produced substitute Consequently almost any import serves both as an input and as an output within the complex matrix of intersectoral flows Changes in import prices will therefore impinge on sectoral activities throughout the economy with ultimate economy-wide implications for the returns to primary factors By aggregating all imports into a single composite good that is treated as distinct from either of the composite final outputs, we have a manageable framework for analyzing the ultimate factoral implications of tariffs

In the next section, I develop and compare the effects of tariffs on factor rewards for the two special cases of a separable and a nonjoint technology In Section III, I make use of the duality between transformation functions and joint cost functions to derive a set of estimating equations which is sufficiently general to test these important restrictions In Section IV, I report the results of an empirical investigation of the model using annual U S postwar data Despite dramatic differences in the determinants of factoral protection which emerge from the two models, both specifications yield the inference that U S tariffs raise real wages and adversely affect the returns to capital These inferences are reversed, however, when we impose additional restrictions on the nonjoint model

II. The Effects of a Tariff on Factor Prices for Separable and Nonjoint Technologies

We assume that the private U S producing sector employs capital, labor, and imported materials to produce tradables (durables and nondurables) and nontradables (services and structures) The aggregate technology is defined by a joint cost function $C(\underline{y}, \underline{w})$ which gives the minimum cost of producing a specified vector of output $\underline{y}$ when the factor price vector is $\underline{w}$ [1] Suppose the producing sector faces exogenous prices for all three commodities, that both final goods are produced at these prices, and that the supplies of primary factors are fixed Under competitive conditions and constant returns to scale, the producing sector achieves an optimal output vector by setting the marginal cost of each output equal to the exogenous output price Since $\partial C(\underline{y}, \underline{w})/\partial y_j$ gives the marginal cost of producing the jth output, we obtain $C_{y_1}(y_1, y_2, w, r, \bar{p}) = \bar{p}_1$, $C_{y_2}(y_1, y_2, w, r, \bar{p}) = \bar{p}_2$, where w and r designate the wage and rental rates, and $\bar{p}$ represents the price of imported materials The partial derivative of the joint cost function with respect to the price of a factor gives the derived demand for that factor, so that $\partial C(\underline{y}, \underline{w})/\partial w_i = x_i(\underline{y}, \underline{w})$ [2] Since primary factors are in fixed supply, the wage and rental rates are implicitly determined by the following factor market equilibrium conditions $C_r(y_1, y_2, w, r, \bar{p}) = \bar{K}$, $C_w(y_1, y_2, w, r, \bar{p}) = \bar{L}$

We now have a system of four simultaneous equations in four unknowns, the final demand vector and the vector of primary-factory prices Consider the effect of the imposition of a tariff on imported materials on the equilibrium vector of factor prices Since all commodity prices are exogenous to the producing sector, the domestic price of the imported input is given by $p = \bar{p}(1 + \tau)$, where τ is the tariff rate The increase in the price of imported inputs will tend to raise the costs of producing final outputs But if these costs are constrained by competitive conditions to equal the unchanged final output prices, adjustments must occur in the prices which can be paid to the primary factors in fixed supply We would also expect a change in the optimum output vector

Without further restrictions on the technology, the solution is a complicated function of all the second-order partial derivatives of the joint cost function [3] It is useful to consider two simplifications of the

---

[1] The joint cost function is developed in detail by Hall (1973) and McFadden (1975) It is defined from the transformation function $T(y, x) = 0$ by $C(y, w) \equiv \text{minimum}_x\, w'x$ subject to $T(y, x) \geq 0$

[2] This is Shephard's lemma See Shephard (1953), Samuelson (1954), Diewert (1971), and McFadden (1975) for details

[3] The solution is easily obtained by totally differentiating the system of four simultaneous equations with respect to small changes in y_1, y_2, w, r, and p An alternative framework for obtaining the comparative static effects on factor prices and outputs involves the dual transformation function Thus, under competitive conditions, the producing sector behaves as if it maximizes national value-added $\bar{p}_1 y_1 + \bar{p}_2 y_2 - pM$ subject to the constraints $T(y_1, y_2, K, L, M) = 0$, $K \leq \bar{K}$, $L \leq \bar{L}$ The Lagrangean for this problem becomes

$$\begin{aligned}\phi(y_1, y_2, K, L, M, \lambda, w, r) \\ \equiv \bar{p}_1 y_1 + \bar{p}_2 y_2 - pM - \lambda T(y_1, y_2, K, L, M) - w(L - \bar{L}) - r(K - \bar{K}),\end{aligned}$$

and the first-order conditions can be differentiated totally with respect to p to obtain the effects on factor rewards This alternative route is more cumbersome, so that the value of duality theory in simplifying the analysis is apparent Burgess (1974*a*) has used the above framework for the one-sector economy technology which have played a prominent role in theoretical and empirical work First, suppose that the technology is functionally separable between inputs and outputs This implies that relative output prices depend only upon output quantities and are independent of factor prices Alternatively, the cost-minimizing input mix is independent of the composition of final output The joint cost function can then be written in the multiplicative form $C(\underline{y}, \underline{w}) = H(\underline{y})\phi(\underline{w})$, where H is linear homogeneous and convex in $\underline{y}$, and ϕ is linear homogeneous and concave in $\underline{w}$ [4] The system of equations appropriate for obtaining the comparative static effects can then be written in matrix form as follows

$$\begin{bmatrix} H_{y_1y_1}\phi(\underline{w}) & H_{y_1y_2}\phi(\underline{w}) & H_{y_1}\phi_w & H_{y_1}\phi_r \\ H_{y_2y_1}\phi(\underline{w}) & H_{y_2y_2}\phi(\underline{w}) & H_{y_2}\phi_w & H_{y_2}\phi_r \\ H_{y_1}\phi_w & H_{y_2}\phi_w & H(\underline{y})\phi_{ww} & H(\underline{y})\phi_{wr} \\ H_{y_1}\phi_r & H_{y_2}\phi_r & H(\underline{y})\phi_{wr} & H(\underline{y})\phi_{rr} \end{bmatrix} \begin{bmatrix} dy_1 \\ dy_2 \\ dw \\ dr \end{bmatrix}$$

$$= \begin{bmatrix} -H_{y_1}\phi_p \\ -H_{y_2}\phi_p \\ -H(\underline{y})\phi_{wp} \\ -H(\underline{y})\phi_{rp} \end{bmatrix} dp$$

The first two equations reduce to a single equation in dw and dr because $H(\underline{y})$ is linear homogeneous [5] If we also premultiply the third and fourth equations by ϕ_r and ϕ_w, respectively, and subtract, we obtain a second equation in dw and dr alone This simplified system of two equations in dw and dr can now be written in matrix form in terms of the first and second partial derivatives of $\phi(w)$ as follows

$$\begin{bmatrix} \phi_w & \phi_r \\ \phi_{ww}\phi_r - \phi_{wr}\phi_w & \phi_{wr}\phi_r - \phi_{rr}\phi_w \end{bmatrix} \begin{bmatrix} dw \\ dr \end{bmatrix} = \begin{bmatrix} -\phi_p \\ \phi_{rp} - \phi_{wp} \end{bmatrix} dp$$

Solving for dw and dr in terms of dp and making use of the Allen (1938) Uzawa (1962) definition for the partial elasticity of substitution $\sigma_{ij} = \phi\phi_{ij}/\phi_i\phi_j$ we obtain

$$\frac{dw}{w} = \frac{p}{w\Delta}\{(\sigma_{wp} - \sigma_{rp}) + (\sigma_{rr} - \sigma_{wr})\phi_r\phi_p\}\frac{dp}{p}$$

$$\frac{dr}{r} = \frac{p}{r\Delta}\{(\sigma_{rp} - \sigma_{wp}) + (\sigma_{ww} - \sigma_{wr})\phi_w\phi_p\}\frac{dp}{p}$$

---

[4] See Hall (1973) for proof

[5] From Euler's Theorem $H_{y_1}y_1 + H_{y_2}y_2 = H(y)$ Therefore, consider an arbitrary change in y_1 and y_2 and differentiate both sides totally to obtain

$$(Hy_1y_1\,dy_1 + Hy_1y_2\,dy_2)y_1 + Hy_1\,dy_1 + (Hy_2y_1\,dy_1 + Hy_2y_2\,dy_2)y_2 + Hy_2\,dy_2 = Hy_1\,dy_1 + Hy_2\,dy_2$$

Consequently, the expressions in brackets must vanish where $\Delta \equiv -\phi_w\phi_r(\sigma_{ww} + \sigma_{rr} - 2\sigma_{wr}) > 0$ since the "unit" cost function ϕ is linear homogeneous and concave in w [6]

These expressions confirm that, while there is a strong presumption that higher import prices reduce both the wage rate and the rental rate, this outcome is not necessary Specifically, if we make the reasonable assumption that capital and labor are substitutes so that $\sigma_{wr} > 0$, the second term in braces for each expression is negative Therefore, the wage (rental) rate will increase only if the elasticity of substitution between imports and capital (labor) is negative and sufficiently large. The fact that only one pair of inputs can be complements rules out the possibility that both factor prices increase with higher tariffs If imported materials are functionally separable from capital and labor, then $\sigma_{wp} = \sigma_{rp}$, and higher tariffs make both types of factor owner worse off [7]

The possibility that imports are closer substitutes for capital than labor, or vice versa, means that tariff policy has an important influence on the distribution of primary-factor income To assess the effect of higher import prices on income distribution, we can derive the following expression

$$\frac{dw}{w} - \frac{dr}{r} = \frac{(\sigma_{wp} - \sigma_{rp})\phi_p}{(\sigma_{ww} + \sigma_{rr} - 2\sigma_{wr})}\phi_w\phi_r\frac{dp}{p},$$

where ϕ_i represents the share of factor payments accruing to the ith factor

To summarize, when the technology is functionally separable between inputs and outputs, the effect of higher import prices on factor prices depends on the relative magnitudes of the partial elasticities of substitution of the "unit" cost function Whether or not labor gains relative to capital depends on whether or not $\sigma_{wp} > \sigma_{rp}$ Thus, for a separable technology, factor demand equations can be specified and income distributional effects can be analyzed as if a single output were produced A diagram will clarify the effects of an import tariff on factor prices for the one-sector (separable) model

Figure 1 gives the locus of wage and rental rates which can be paid per unit of the primary factors under competitive conditions when one unit of final output sells for a dollar and imported materials can be purchased

---

[6] Since the unit cost function ϕ is linear homogeneous and concave in w, then its Hessian matrix is negative semidefinite Consequently, the matrix of Allen-Uzawa partial elasticities of substitution is also negative semidefinite Therefore,

$$\sigma_{ww} + \sigma_{rr} - 2\sigma_{wr} = (1 \quad -1)\begin{pmatrix}\sigma_{ww} & \sigma_{rw} \\ \sigma_{wr} & \sigma_{rr}\end{pmatrix}\begin{pmatrix}1 \\ -1\end{pmatrix} < 0.$$

[7] This result follows easily since the expressions in the text simplify to $dw/w = (p/w\Delta)(\sigma_{rr} - \sigma_{wr})\phi_r\phi_p\ (dp/p)$, and $dr/r = (p/r\Delta)(\sigma_{ww} - \sigma_{wr})\phi_w\phi_p\ (dp/p)$, which are both negative whenever import prices rise

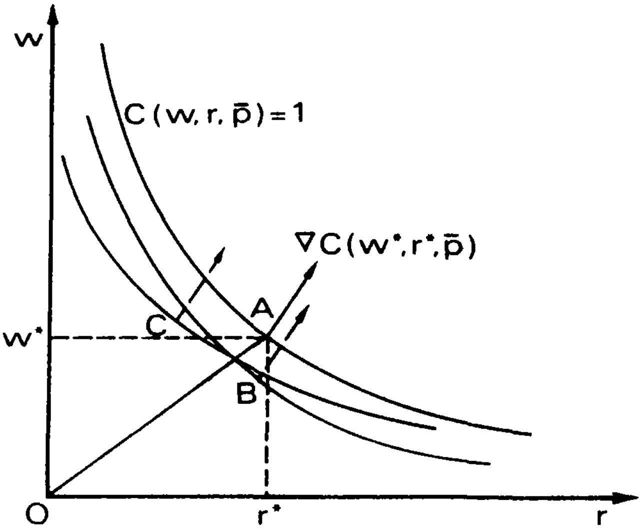


FIG 1

at price $\bar{p}$[8] By Shephard's lemma, the competitive equilibrium factor-price vector will be given by the point A on the locus where the gradient vector coincides with the economy's fixed-endowment vector Now suppose p is increased through the imposition of a tariff The new locus of factor prices which can be paid when final output continues to sell for a dollar must be everywhere inside the old locus[9] The new equilibrium factor prices will be B, which represents the unique point where the gradient vector has the same direction as before This corresponds to the full-employment condition imposed on the unchanged supply of primary factors, capital and labor

The change in factor prices clearly depends on the manner in which higher import prices shift the isocost locus A neutral shift occurs when $\sigma_{wp} = \sigma_{rp}$, implying that capital and labor are equivalent substitutes for imported materials If $\sigma_{wp} > \sigma_{rp}$, then the wage rate declines less than the rental rate with higher tariffs (the wage/rental rate increases) The new locus has a flatter slope along the ray OA If $\sigma_{wp} < \sigma_{rp}$, then the wage rate declines more than the rental rate, and the new locus has a steeper slope along the ray OA Higher tariffs might raise the real reward for one factor provided it is complementary to imports For example, if imports and capital were complements so that $\sigma_{rp} < 0$, then the twist in the locus might be severe enough that equilibrium is restored at C where wages have increased.

---

[8] The locus of factor prices w, r is quasi-concave for a given value of p because the value-added function exhibits all the properties of a well-behaved technology, assuming that the overall technology for final output is well behaved See Woodland (1973) and Diewert (1974) for further details

[9] Since a unit of final output sells for the same price, while the price of imported materials has increased, less can be paid per unit of capital and labor in order that unit costs remain unchanged

The two-sector nonjoint production model is the cornerstone of a great deal of theoretical work[10] Theorists have become accustomed to treating multiple-output firms or economies as consisting of distinct sectors each with separate production functions The cost of producing any vector of final output is just the sum of the costs of producing each output separately Provided the technology exhibits constant returns to scale, the joint cost function can then be written $C(\underline{y}, \underline{w}) = y_1C^1(\underline{w}) + y_2C^2(\underline{w})$, where $C^i(\underline{w})$ is the unit cost function for the ith sector[11] Under this specification, output prices depend only on factor prices, and aggregate factor demands are the sum of individual sector demands This means that the matrix system for determining the comparative static effects on factor rewards and outputs can be written

$$\begin{bmatrix} 0 & 0 & C^1_w & C^1_r \\ 0 & 0 & C^2_w & C^2_r \\ C^1_w & C^2_w & y_1C^1_{ww} + y_2C^2_{ww} & y_1C^1_{wr} + y_2C^2_{wr} \\ C^1_r & C^2_r & y_1C^1_{rw} + y_2C^2_{rw} & y_1C^1_{rr} + y_2C^2_{rr} \end{bmatrix} \begin{bmatrix} dy_1 \\ dy_2 \\ dw \\ dr \end{bmatrix} = \begin{bmatrix} -C^1_p \\ -C^2_p \\ -y_1C^1_{wp} - y_2C^2_{wp} \\ -y_1C^1_{rp} - y_2C^2_{rp} \end{bmatrix} dp$$

The subsystem of the first two equations yields a solution for the effects on factor prices independent of the solution for outputs If we define θ_{ij} as the cost share of the ith factor in the jth sector, we can express the percentage changes in factor prices resulting from a given percentage change in import prices as follows[12]

$$\frac{dw}{w} = \left\{\frac{\theta_{K1}\theta_{M2} - \theta_{K2}\theta_{M1}}{|\theta|}\right\}\frac{dp}{p}$$

$$\frac{dr}{r} = \left\{\frac{\theta_{L2}\theta_{M1} - \theta_{L1}\theta_{M2}}{|\theta|}\right\}\frac{dp}{p}$$

---

[10] For examples of its use in international trade, the reader is referred to Stolper and Samuelson (1941), Jones (1971), and Khang (1973) Johnson (1971) provides an excellent survey of the two-sector model with applications to problems of tax incidence and unionization

[11] See Hall (1973) for proof

[12] The two equations can be written in the form $C^i_w\,dw + C^i_r\,dr = -C^i_p\,dp$ ($i = 1, 2$) We can then convert the system into a form where the coefficients are factor shares and the unknowns are percentage rates of change

$$\frac{wC^i_w}{y_i}\frac{dw}{w} + \frac{rC^i_r}{y_i}\frac{dr}{r} = \frac{-pC^i_p}{y_i}\frac{dp}{p}$$

($i = 1, 2$), which is simply $\theta_{Li}\,dw/w + \theta_{Ki}\,dr/r = -\theta_{mi}\,dp/p$ ($i = 1, 2$)

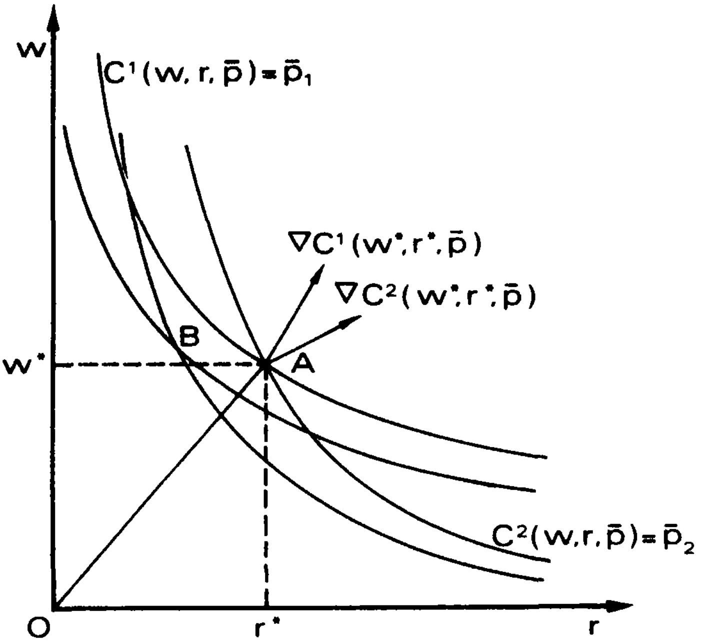

w
$C^1(w, r, \bar{p}) = \bar{p}_1$
$\nabla C^1(w^*, r^*, \bar{p})$
$\nabla C^2(w^*, r^*, \bar{p})$
B
A
w^*
$C^2(w, r, \bar{p}) = \bar{p}_2$
O
r^*
r

FIG 2

where $|\theta| = \theta_{L1}\theta_{K2} - \theta_{L2}\theta_{K1} > 0$, provided the value-added process for the first sector is labor intensive at the initial factor prices [13]

For the nonjoint model, provided the tariff revenue is distributed in lump-sum fashion, the effects of higher tariffs on factor prices depend only on the primary-factor intensities and import cost shares for each sector To clarify this, consider figure 2, which gives the isocost loci in factor-price space for each sector at the initial set of commodity prices With free mobility of factors between sectors and perfectly competitive conditions, the unique set of factor prices is given by the intersection of these loci Since the gradient vectors to each isocost locus at this point represent the unit factor demands for each sector, the equilibrium output vector is given by the coordinates of the endowment vector in the basis of unit factor demands

Assume that imported inputs increase in price as a consequence of higher tariffs, thereby shifting both isocost loci inward The extent of the inward shift of the isocost loci along the ray *OA* depends only on the share of imports in the cost of final output Since the new equilibrium set of factor prices must be an intersection of the new isocost loci, the point will lie above or below the ray, depending first on the share of imports in each sector and second on the relative primary-factor intensities of each sector The geometric analysis confirms what the algebra has already indicated; for small changes in the price of imported inputs, initial primary-factor intensities and import cost shares are the sole determinants of changes in factor prices Whether the isocost loci twist in a biased fashion depends, as before, upon sectoral differences in elasticities of substitution between imports on the one hand and capital and labor on the other But for small changes in import prices, this aspect is irrelevant in the determination of factor prices Thus, in the two-sector model, both sectors may exhibit complementarity between capital and imports (thereby twisting both loci so that the contours are flatter along *OA*), and yet this may imply that the rental rate rises and the wage rate falls In the one-sector model, this complementarity was a necessary, but not sufficient, condition for the wage rate to rise and the rental rate to fall[14]

---

[13] The condition $|\theta| > 0$ iff $\theta_{L1}\theta_{K2} > \theta_{L2}\theta_{K1}$ requires that $\theta_{L1}/\theta_{K1} > \theta_{L2}/\theta_{K2}$, which says $w_1L_1/r_1K_1 > w_2L_2/r_2K_2$ If we assume factor mobility and no market distortions, so that factors earn the same rates of remuneration in each sector, then this latter condition implies $L_1/K_1 > L_2/K_2$

In concluding this section, it is useful to consider the special case where imported materials enter only one sector Geometrically, it is easy to see that since only one isocost locus shifts inward, one group of factor owners will necessarily be better off with higher tariffs on imported materials More precisely, from the algebra it follows that if the sector requiring no imported inputs is labor (capital) intensive, the wage (rental) rate increases with higher tariffs no matter what the relative magnitudes of the elasticities of substitution happen to be This provides a dramatic contrast to the determinants for the one-sector (separable) model[15]

III. Empirical Implementation

Functional Form

We assume that the technology of the producing sector for the private U S economy can be represented by the following generalized linear–

---

[14] To show that both factors cannot be better off we need only show that both numerators cannot be positive when $|\theta| > 0$ Suppose, to the contrary, that both numerators are positive so that $\theta_{K1}/\theta_{m1} > \theta_{K2}/\theta_{m2}$ and $\theta_{m1}/\theta_{L1} > \theta_{m2}/\theta_{L2}$. Together these conditions imply that $\theta_{K1}/\theta_{L1} > \theta_{K2}/\theta_{L2}$, which contradicts the assumption that $|\theta| > 0$ On the other hand, both expressions can be negative, indicating that both factor owners might be worse off with tariff protection Thus, if $\theta_{L1}/\theta_{m1} > \theta_{L2}/\theta_{m2}$ while $\theta_{m1}/\theta_{K1} > \theta_{m2}/\theta_{K2}$, this will guarantee that $|\theta| > 0$, although it must be emphasized that $|\theta| > 0$ can hold without both expressions being negative

[15] The effect of higher tariffs on the distribution of income between capital and labor for the two-sector nonjoint model is given by

$$dw/w - dr/r = \left\{\frac{\theta_{m2} - \theta_{m1}}{|\theta|}\right\} dp/p$$

If the first sector is labor intensive, the share of factor income accruing to labor will increase provided the share of imported materials is smaller for this sector.

generalized Leontief (GLGL) joint cost function in two outputs and three inputs

$$C(\underline{y}, \underline{w}) = \sum_{i}^{2} \sum_{j}^{2} \sum_{r}^{3} \sum_{s}^{3} \alpha_{ijrs}\sqrt{y_i y_j}\sqrt{w_r w_s}$$

The *GLGL* functional form is sufficiently flexible to provide a second-order approximation to an arbitrary cost function in a small neighborhood about the point of evaluation[16] It also enables us to test the two important restrictions of separability and nonjointness which have frequently been adopted as alternative maintained hypotheses in most previous empirical work Separability means that the joint cost function can be written in the multiplicative form

$$C(\underline{y}, \underline{w}) = \left(\sum_{i}^{2} \sum_{j}^{2} \alpha_{ij}\sqrt{y_i y_j}\right)\left(\sum_{r}^{3} \sum_{s}^{3} \beta_{rs}\sqrt{w_r w_s}\right),$$

where the second bracketed expression can be viewed as the "unit" input cost function Separability gives rise to the restrictions $\alpha_{iijj}\alpha_{rrss} = \alpha_{iirr}\alpha_{jjss}$ for all i, j, r, s Nonjointness means that the joint cost function can be written in the form

$$C(\underline{y}, \underline{w}) = \sum_{i}^{2} y_i \left(\sum_{r}^{3} \sum_{s}^{3} \alpha_{iirs}\sqrt{w_r w_s}\right),$$

where the bracketed expression represents the unit cost function for the ith sector Nonjointness gives rise to the restrictions $\alpha_{ijrs} = 0$ unless $i = j$ Since the function admits separability and nonjointness as special cases, we have an analytical framework which is appropriate for testing the reasonableness of the two competing models for explaining the income distributional effects of tariffs

The flexibility of the GLGL form is evident from the fact that it permits an arbitrary set of Allen-Uzawa partial elasticities of substitution between pairs of inputs at a given set of factor prices and output quantities Contrary to the multifactor generalizations of the Cobb-Douglas and constant elasticity of substitution (CES) forms, separability restrictions among the inputs are not imposed a priori, and complementarity is possible between a pair of factors (see Arrow et al [1961]) The partial elasticity of substitution between factors r and s has the following general representation in terms of the joint cost function $\sigma_{rs} = CC_{rs}/C_rC_s$ It has the following specific representation in terms of the parameters of the GLGL joint cost function:

---

[16] Principles of duality assure us of the existence of a well-behaved technology associating inputs to outputs, which is implied by any given joint cost function provided it exhibits certain characteristics (see McFadden 1975) The generalized linear-generalized Leontief joint cost function was developed by Hall (1973) from the generalized Leontief cost function for a single output proposed by Diewert (1971) The first-order terms are suppressed because of our maintained hypothesis of constant returns to scale

$$\sigma_{rs} = \frac{(\sum_i \sum_j \sum_r \sum_s \alpha_{ijrs}\sqrt{y_i y_j}\sqrt{w_r w_s})(\sum_i \sum_j \alpha_{ijrs}\sqrt{y_i y_j / w_r w_s})}{(\sum_i \sum_j \sum_r \alpha_{ijrs}\sqrt{y_i y_j}\sqrt{w_r / w_s})(\sum_i \sum_j \sum_s \alpha_{ijrs}\sqrt{y_i y_j}\sqrt{w_s / w_r})}$$

This measures the normalized response of the demand for the rth input to a change in the price of the sth input when all other input and output prices are unchanged and when all inputs and outputs are allowed to vary It is therefore equivalent to the comparative static effect of a change in one input price for a price-taking firm except for the normalization which ensures that σ_{rs} and σ_{sr} are equal [17]

Data

In the *National Income and Product Accounts of the United States* (U S Department of Commerce 1966, tables 1 4, 1 6), merchandise imports are subtracted from the various categories of final output (durables and nondurables) in order to arrive at measures of value-added outputs from the standpoint of the private domestic economy This procedure reflects the implicit assumption that imports are final goods requiring no further domestic processing I have argued that a more appropriate treatment of imports from the standpoint of assessing factoral protection is as an additional input along with domestically supplied capital and labor services in the production process for final goods In this paper, I have decomposed final output into durables and nondurables on the one hand and nongovernment services and structures on the other I add merchandise imports back into the national-accounts measure of durables and nondurables output to obtain measures of final output in current and constant dollars Prices and quantities of durables, nondurables, nongovernment services and structures are measured from the producer's point of view and therefore take into account producer taxes and subsidies I follow Christensen and Jorgenson (1970) and include their measure of the imputed value of the services of consumer and producer durables, while excluding all services supplied by governments [18]

Prices and quantities of capital and labor services are also taken from Christensen and Jorgenson (1970) and reflect producer taxes on inputs These authors construct a quantity index for labor input based upon private domestic persons engaged adjusted for hours worked and changes in the level of educational attainment. The capital-service price and quantity data are constructed by Divisia aggregation from data on the capital services provided by seven distinct capital goods, including land All data are annual for the period 1948–69

---

[17] See Diewert (1974) for further details

[18] See Christensen and Jorgenson (1970) for a detailed discussion of the production account These authors divide output into durables and structures on the one hand and nondurables and services on the other A detailed account of the data and the procedure used in the disaggregation of output used in this paper is available from me on request

Jorgenson and Griliches (1967) have shown that most but not all of the growth in aggregate real output can be explained by the growth in factor input when secular changes in labor skill and the composition of capital goods is taken into account In this paper I adopt as a maintained hypothesis that all disembodied technical change is Hicks neutral In order to derive an index of neutral technical change, I compute measures of aggregate input and output as Divisia indexes of individual inputs and outputs where the weights are value shares The index of disembodied technical change is obtained by dividing the index of aggregate output by the index of aggregate input This index is then multiplied into the initial measures of inputs and output prices in order to obtain measures which are adjusted for disembodied technical change

Estimation Procedure

I propose to use multivariate regression techniques to obtain maximum-likelihood estimates of the set of factor demand and output price equations obtained from the GLGL joint cost function If we differentiate the cost function with respect to each factor price, we obtain three estimating equations for the cost-minimizing factor demands The following demand equation is obtained for the first factor

$$\begin{aligned}\partial C(\underline{y}, \underline{w})/\partial r = K = {} & \alpha_{11KK}y_1 + \alpha_{22KK}y_2 + \alpha_{11KL}y_1\sqrt{w/r} \\ & + \alpha_{11KM}y_1\sqrt{p/r} + \alpha_{22KL}y_2\sqrt{w/r} + \alpha_{22KM}y_2\sqrt{p/r} \\ & + 2\alpha_{12KK}\sqrt{y_1 y_2} + 2\alpha_{12KL}\sqrt{y_1 y_2}\sqrt{w/r} \\ & + 2\alpha_{12KM}\sqrt{y_1 y_2}\sqrt{p/r}\end{aligned}$$

If we differentiate the joint cost function with respect to each output, we obtain expressions for the marginal costs of producing each output Under competitive conditions and constant returns to scale, these expressions can be identified with the respective output prices The following output-price equation is obtained for the first output

$$\begin{aligned}\partial C(\underline{y}, \underline{w})/\partial y_1 = p_1 = {} & \alpha_{11KK}r + \alpha_{11LL}w + \alpha_{11MM}p + 2\alpha_{11KL}\sqrt{r\cdot w} \\ & + 2\alpha_{11KM}\sqrt{r\ p} + 2\alpha_{11LM}\sqrt{w\cdot p} + \alpha_{12KK}r\sqrt{y_2/y_1} \\ & + \alpha_{12LL}w\sqrt{y_2/y_1} + 2\alpha_{12MM}p\sqrt{y_2/y_1} \\ & + 2\alpha_{12KL}\sqrt{r\ w}\sqrt{y_2/y_1} + 2\alpha_{12KM}\sqrt{r\ p}\sqrt{y_2/y_1} \\ & + 2\alpha_{12LM}\sqrt{w\ p}\sqrt{y_2/y_1}\end{aligned}$$

Because the five estimating equations contain parameters in common, consistent estimates will be obtained only if the individual equations are stacked and estimated as a single equation. Suppose we adopt a stochastic specification which reflects (1) errors in setting price equal to marginal cost and (2) errors in attaining the cost-minimizing factor inputs, by adding the classical disturbance term to each equation. The overall variance-covariance matrix of disturbances will be singular, however, in view of the accounting identity that income equals output each period. I propose to drop one factor demand equation and use the residuals from the stacked ordinary least squares (OLS) regression to obtain estimates of the "true" variance-covariance matrix of disturbances. Disturbances drawn contemporaneously from each factor demand and output-price equation are likely to be correlated. This follows perhaps because the underemployment of one factor may imply a systematic overemployment of another factor, and the understatement of the marginal cost of one output may imply a systematic overstatement of the marginal cost of the other output If we transform the variables so that the transformed matrix of residuals becomes an identity matrix, and if we perform OLS on the transformed variables and iterate until the matrix of disturbances converges to an identity matrix, our iterative Zellner efficient parameter estimates will possess maximum-likelihood properties and be invariant with respect to the choice of omitted equation [19]

IV. Empirical Results

The GLGL form imposes linear homogeneity in factor prices and outputs a priori I have also adopted Hicks neutral disembodied technical change as a maintained hypothesis The remaining important well-behaved properties of the technology are testable hypotheses, they provide some indication of the wisdom of using the essentially microeconomic theory of optimizing behavior at the aggregate level The joint cost function must be nondecreasing and concave in factor prices and nondecreasing and convex in outputs These properties will obtain provided (1) $\partial C/\partial w_r \geq 0$ for all r while $\partial^2 C/\partial w_r\, \partial w_s$ is a negative semidefinite matrix, and (2) $\partial C/\partial y_i \geq 0$ for all i while $\partial^2 C/\partial y_i\, \partial y_j$ is a positive semidefinite matrix at all data points The derived factor demand and output-price equations contain parameters in common These parametric restrictions must be imposed in order for the estimating equations to be consistent with the underlying theory of optimizing behavior While one can view the imposition of these "symmetry" restrictions as the efficient use of valuable prior information, a comparison of the explanatory power of the model with and without these restrictions imposed provides some indication of the validity of the underlying theory. Failure of either the symmetry restrictions to pass the conventional statistical tests or the monotonicity and convexity properties to hold brings into question the wisdom of a variety of standard neoclassical assumptions for aggregative analysis· two examples are the complete omission of lags in optimizing behavior as reflected in the assumption that relative output prices equal marginal rates of transformation, and the assumption of a sufficiently malleable capital stock and labor force so that internal efficiency in resource allocation is achieved [20] One might also question the validity of the *maintained hypotheses of constant returns to scale and neutral dis*embodied technical change

---

[19] See Zellner (1962), Kmenta and Gilbert (1968), and Chow and Fair (1971)

The first step in the estimation procedure is to drop the import demand equation and combine the capital and labor demand equations with the two output-price equations to form a single estimating equation with the across-equation symmetry restrictions imposed Estimation of this single stacked equation by ordinary least squares will yield parameter estimates which are not invariant to the choice of omitted equation To achieve invariant estimates and to gain efficiency, we construct an estimate of the variance-covariance matrix of contemporaneous disturbances from the residuals of the OLS regression and proceed to minimize the determinant of this matrix using Zellner's iterative Aitken estimation procedure [21] The constrained iterative Zellner efficient (IZEF) estimates of the 18 free parameters of the GLGL joint cost function based upon 88 observations appear in column 1 of table 1

I have already suggested the dual role of the symmetry restrictions on the one hand as a source of additional prior information and on the other hand as a test of the validity of the specification Without imposing these restrictions, we have a set of four separate estimating equations, each based upon 22 observations, to explain movements in output prices and factor demands There are nine free parameters in each factor demand equation and 12 free parameters in each output-price equation The sum of squared residuals under the null hypothesis of symmetry can be compared to the sum of squared residuals for the unconstrained equations The calculated F-statistic is 1 99, while the critical $F(24, 46)$ is 1 75 at the 5-percent level and 2 22 at the 1-percent level I conclude that there is rather strong evidence against symmetry, and this opens up the distinct possibility that one or more of the maintained hypotheses of our model are invalid for purposes of aggregative analysis

[20] See Mayer (1974) for further discussion of the short-run implications of sector specificity of factor services in the two-sector nonjoint model

[21] Output prices and quantities and input prices and quantities are undoubtedly simultaneously determined at the level of the aggregate economy Therefore, our IZEF estimates probably exhibit some simultaneous equation bias Preliminary experimentation with iterative three stages least squares (I3SLS) estimation indicated very little change in the point estimates or their standard errors but gave rise to substantial increases in computational costs

TABLE 1

Iterative Zellner-efficient Estimates of Generalized Linear–Generalized Leontief Joint Cost Function

	General	Nonjoint	Separable (Implied Estimates)		Separable (Direct Estimates)
α_{11KK}	−120.009	2.2345	0.0984	α_{11}	0.0984
	(51.697)	(0.9051)	(0.0386)		(0.0386)
α_{11LL}	0.8768	1.4632	0.1157	α_{22}	0.0965
	(42.048)	(0.5651)	(0.0465)		(0.0363)
α_{11MM}	−51.151	−0.5752	−0.0288	α_{12}	0.0275
	(32.569)	(0.4235)	(0.0139)		(0.0283)
α_{11KL}	35.076	−1.3901	0.2468	β_{rw}	2.5093
	(31.870)	(0.5944)	(0.0772)		(1.2784)
α_{11KM}	94.762	−0.3688	−0.0421	β_{rp}	−0.4277
	(37.417)	(0.5034)	(0.0169)		(0.1399)
α_{11LM}	−40.919	0.7043	0.0969	β_{wp}	0.9854
	(24.833)	(0.3711)	(0.0245)		(0.3703)
α_{22KK}	−101.756	−1.2897	0.0965	β_{ww}	1.1731
	(39.514)	(0.6941)	(0.0363)		(0.3770)
α_{22LL}	−5.3768	−1.0037	0.1132	β_{pp}	−0.2928
	(31.659)	(0.4398)	(0.0420)		(0.2085)
α_{22MM}	−44.309	0.4731	−0.0283	β_{rr}	1.000
	(25.069)	(0.3268)	(0.0128)		
α_{22KL}	34.722	1.6220	0.2422		
	(24.046)	(0.4535)	(0.0637)		
α_{22KM}	74.784	−0.0319	−0.0413		
	(28.952)	(0.3908)	(0.0163)		
α_{22LM}	−33.475	−0.1855	0.0951		
	(18.417)	(0.2763)	(0.0204)		
α_{12KK}	−110.916		0.0275		
	(45.165)		(0.0283)		
α_{12KL}	−34.886		0.0322		
	(27.653)		(0.0284)		
α_{12KM}	−84.379		0.0689		
	(32.910)		(0.0554)		
α_{12LL}	2.7739		−0.0118		
	(36.475)		(0.0126)		
α_{12LM}	37.234		0.0271		
	(21.378)		(0.0247)		
α_{12MM}	50.785		−0.0080		
	(28.575)		(0.0074)		
D-W	1.8345	1.5164	0.5598		
SSR	3.9978	5.0256	19.0820		
SE	(0.23838)	(0.25715)	(0.48839)		
OBS	88	88	88		

Note.—Symmetry conditions imposed, standard errors in parentheses. SSR = sum of squared residuals, OBS = observations.

Whether the symmetry restrictions should be viewed as valid prior information or as a test of model specification perhaps depends to some extent on whether the model with the restrictions imposed meets the minimum requirements for a well-behaved technology. Using the point estimates from column 1, we find that the joint cost function is increasing and positive semidefinite in outputs for all periods and increasing and negative semidefinite in factor prices for most periods of the postwar

TABLE 2

IMPLIED POINT ESTIMATES OF OWN-PRICE ELASTICITIES OF DEMAND AND ALLEN-UZAWA PARTIAL ELASTICITIES OF SUBSTITUTION

	η_K	η_L	η_M	σ_{wr}	σ_{rp}	σ_{wp}
1948	−8.061	−5.242	−1.056	17.605	−0.467	2.528
1949	−3.135	−4.214	−0.174	10.662	−9.861	5.746
1950	−3.927	−4.535	−0.269	11.469	−9.613	6.013
1951	−5.015	−4.812	−0.507	13.484	−6.867	4.829
1952	−4.451	−4.543	−0.398	12.772	−7.173	4.705
1953	−4.555	−4.398	−0.425	12.879	−6.249	4.202
1954	−0.229	−3.260	−0.195	6.158	−13.626	7.448
1955	−1.415	−3.667	−0.039	8.008	−12.597	6.861
1956	−0.564	−3.276	−0.187	6.709	−12.045	6.547
1957	0.031	−2.905	−0.430	5.298	−11.563	6.603
1958	0.313	−2.171	−1.579	2.700	−9.183	7.506
1959	0.231	−2.464	−0.864	3.719	−10.688	7.137
1960	0.055	−2.634	−0.584	4.549	−10.473	6.396
1961	0.224	−2.292	−1.002	3.366	−9.194	6.538
1962	−0.127	−2.623	−0.526	4.571	−10.642	6.574
1963	−0.373	−2.754	−0.388	5.175	−10.612	6.308
1964	−0.919	−2.992	−0.264	6.355	−10.154	5.782
1965	−1.265	−3.138	−0.204	6.812	−10.457	5.974
1966	−2.146	−3.402	−0.201	8.214	−9.209	5.318
1967	−1.387	−3.074	−0.248	7.105	−9.207	5.209
1968	−2.016	−3.189	−0.264	8.221	−7.967	4.500
1969	−2.056	−3.156	−0.274	8.227	−7.760	4.389

sample. A clearer indication of the characteristics of the technology is obtained from table 2, which presents the estimates of the Allen-Uzawa partial elasticities of substitution and own price elasticities of demand for each period. Thus, despite the large standard errors of 12 of the 18 free parameter estimates, which results from the substantial multicollinearity of annual postwar data, the point estimates confirm that the joint cost function is essentially well behaved.[22] Perhaps the most interesting feature of table 2 is the strong indication that capital services and imports are complementary inputs. In view of the discussion in Section II, we cannot draw any inferences about the effects of tariffs on factor rewards from this evidence unless we know whether the technology is separable or nonjoint. I now proceed to examine the validity of these alternative specifications and the inferences which emerge from them.

Separability

For the separable model, changes in input prices have no differential impact on the optimum output combination at given output prices, an ncrease in tariffs will reduce imports, and production of both types of output will contract in proportion Conceptually, inputs are combined to produce an aggregate input which is then allocated between the production of outputs, so that the value of its marginal product at the given output prices is equalized among activities Since the optimum input combination depends only upon factor prices, the imposition of a tariff will affect the prices of domestically supplied primary factors in a manner which depends only upon the partial elasticities of substitution, as discussed in Section II

---

[22] The joint cost function fails to satisfy the monotonicity and convexity conditions for the period 1957–61, during which the own-price elasticity of demand for capital services is positive. One possible explanation for this is the possibility of lags in optimizing behavior. We feel that a useful examination of lags in optimizing behavior must begin with quarterly data on factor demands.

We have seen that the separable technology implies that the joint cost function can be written as the product of what we can call the unit input cost function $\phi(\underline{w})$ and the unit output frontier $H(\underline{y})$ The unit input cost function is identical in form to Diewert's (1971) generalized Leontief cost function for a single output and contains six parameters when there are three inputs The unit output frontier has a generalized linear relationship between outputs and contains three parameters when there are two outputs However, only eight of the nine parameters are free for estimation purposes because the parameters of the joint cost function are the products in pairs of the parameters of the input and output functions There is no natural measure of a unit of aggregate input, so it makes no difference which of the nine parameters we constrain to be unity I choose to normalize on the coefficient of the price of capital services in the unit cost function, leaving us with eight free parameters The nonlinear estimation procedure is to begin with plausible initial values for these parameters which determine the 18 parameters of the joint cost function through the ystem of nonlinear constraints) and to iterate on these parameters until he residual sum of squares is minimized

Column 3 of table 1 presents the implied estimates of the 18 parameters orresponding to the separable model[23] Despite the nonlinearity of the estrictions, the F-statistic provides a reasonably accurate test of the estrictions Adopting the constrained GLGL form as the maintained ypothesis, the calculated value of F with 10 and 70 df is 26 49, which s substantially above the 1-percent critical value of 2 59 Separability is herefore decisively rejected as a specification of the technology of the ggregate producing sector I conclude that postwar movements in the elative prices of the two types of output cannot adequately be explained n terms of changes in the composition of output alone, changes in factor rices are also important determinants of changes in relative output rices

---

[23] It should not be surprising to find substantial changes in the point estimates of the arameters when moving from the unrestricted GLGL form to either the nonjoint or eparable models In general, the individual parameters have no clear behavioral nterpretation The interesting economic parameters such as price and substitution lasticities are complicated functions of these parameters, and these functions change when the restrictions of separability and nonjointness are imposed

Despite the convincing evidence against the separable model, it is important to examine the implications of this specification for the income distributional effects of tariffs The separable model plays a prominent role in macroeconomic models of income determination and growth, since it is the implicit maintained hypothesis of economists who apply the one-sector model to situations where several types of output exist (for example, consumption and investment goods or tradables and non-tradables) The easiest way to focus on the specific characteristics of the separable technology is to examine the direct estimates of the unit input cost function and the unit output frontier, which I have presented in column 4 of table 1 The first observation to be made is that, with one minor exception, the parameter estimates are highly significant and yield all the properties of a well-behaved technology Thus, the unit input cost function

$$\phi(\underline{w}) = \sum_{r} \sum_{s} \beta_{rs} \sqrt{w_r w_s}$$

is increasing and concave in factor prices for every period, and the unit output frontier

$$H(\underline{y}) = \sum_{i} \sum_{j} \alpha_{ij} \sqrt{y_i y_j}$$

is increasing but not convex in outputs, since the point estimate of the interaction term α_{12} is positive However, we clearly cannot reject the convexity property because of the high standard error attached to this estimate

What are the implications of higher tariffs for factor rewards and income distribution arising from the separable model? Recalling that $\sigma_{rs} = \phi\phi_{rs}/\phi_r\phi_s$, it is easy to compute the following implied point estimates for the Allen-Uzawa partial elasticities of substitution for the base period 1958 $\sigma_{rw} = 71$, $\sigma_{rp} = -2\ 12$, $\sigma_{wp} = 3\ 18$ In Section II I showed that if $\sigma_{wp} > \sigma_{rp}$, then real income accruing to capital is reduced and income is distributed from capital to labor with higher tariffs Furthermore, laborers will actually be made better off with higher tariffs, provided $\sigma_{rp} - \sigma_{wp} - (\sigma_{rr} - \sigma_{wr})\phi_r\phi_p < 0$ Making use of the definition of the Allen-Uzawa partial elasticity of substitution, this condition can be written in terms of the first and second partial derivatives of ϕ as $\phi_{rp}\phi_w - \phi_{wp}\phi_r - (\phi_{rr}\phi_w - \phi_{wr}\phi_r)\phi_p^2 < 0$ Using the point estimates, this expression takes the value of $-1\ 86$ for 1958 The precise standard error for this statistic has not been calculated because of the nonlinearity of the expression in the parameters However, we have tentative evidence from the separable model that higher tariffs actually make labor better off[24]

---

[24] We see that $\sigma_{rp} = \phi\phi_{rp}/\phi_r\phi_p$ Therefore, provided the technology is well behaved, sign σ_{rp} = sign β_{rp} The evidence regarding complementarity between capital and imports is decisive because of the low standard error attached to $\hat{\beta}_{rp}$ Furthermore, we decisively reject the hypothesis that the input function is separable with respect to a partitioning between capital and labor on the one hand and imports on the other (i.e , the hypothesis that $\sigma_{wp} = \sigma_{rp}$)

Nonjointness

The unrestricted GLGL form is sufficiently general to explain changes in output prices and factor demands in terms of changes in factor prices and changes in the composition of output The special case of nonjoint production implies that output prices are determined by factor prices alone, and that aggregate factor demands are just the sum of individual sectoral demands The joint cost of producing any vector of output under these conditions can be written as the sum of unit costs for each sector weighted by the respective levels of sector output $C(\underline{y}, \underline{w}) = y_1C^1(\underline{w}) + y_2C^2(\underline{w})$, where each unit cost function has a generalized Leontief form

Column 2 of table 1 presents the estimates of the 12 free parameters of the nonjoint model Adopting the GLGL form with the symmetry conditions as the maintained hypothesis, the calculated value of the F-statistic with 6 and 70 df is 3 07, which is precisely equal to the critical value of F at the 1-percent level The evidence therefore indicates that changes in the relative prices of the composite outputs we distinguish cannot be explained adequately by changes in factor prices alone, nevertheless, the nonjoint model performs better in explaining the data than the separable model

We should not be too surprised by the statistical evidence against the nonjoint model The disaggregation of output into two composite types which are assumed to be produced separately using primary factors is a convenient analytical device, and it is likely to be an appropriate empirical assumption the more closely the production processes for the diverse individual outputs being aggregated to form the composite output are similar in their primary-factor intensities Strictly speaking, the existence of consistent measures of composite outputs from the standpoint of production theory requires that the separate production processes be identical for each individual output being aggregated to form the composite output [25] Viewed from this perspective, our choice of composite outputs is unlikely to meet these requirements or even approximate them The two types of output we aggregate to form the nontraded-goods category, namely nongovernment services and structures, typically exhibit substantially different primary-factor intensities, many services tend to be highly labor intensive, while structures require large amounts of physical capital The aggregation of durables and nondurables to form the traded-goods sector is perhaps less susceptible to this criticism I conclude that, while a more detailed view of the economy and perhaps even a different two-output aggregation criterion might support the validity of the nonjoint model, there may be serious specification errors involved in using the nonjoint model based upon a two-output categorization of the economy's output into tradables and nontradables

---

[25] More accurately, the production functions must differ by at most a neutral scale factor

Despite the statistical evidence against this particular nonjoint specification, a detailed examination of the specific characteristics of the nonjoint model is warranted for two reasons· because the evidence against it is not nearly as overwhelming as the evidence against separability, and also because the distinction between the nonjoint production model based upon tradables and nontradables is frequently employed by theorists in discussing issues such as devaluation in international trade I now focus on the implications of the parameter estimates of the nonjoint model for the income distributional effects of tariffs. The point estimates indicate that nontraded goods (services and structures) are labor intensive since $|\hat{\theta}| = \hat{\theta}_{L1}\hat{\theta}_{K2} - \hat{\theta}_{K1}\hat{\theta}_{L2} > 0$[26] In addition, services and structures have a lower import requirement per dollar of final output From the analysis of Section II I conclude that higher tariffs distribute real income from capital to labor The reasoning behind this conclusion is that higher tariffs tend to raise unit costs more substantially in the traded-goods sectors because of their higher import requirement But since these sectors tend to be capital intensive and since final output prices are fixed by assumption, then the change in primary-factor prices needed to restore unit costs involves a reduction in rental rates and an increase in wage rates

The nonjoint model is less successful in explaining the substitution possibilities among inputs for each output category Using the point estimates, neither sector exhibits all the properties of a well-behaved technology The cost functions fail to be concave in factor prices for either sector, because two of three α_{iirs} terms are negative ($i = 1, 2, r \neq s$), although we cannot reject concavity for either sector[27] The cost function for nontraded goods also fails to be a monotonic increasing function of factor prices because of the implied negative import demand requirement, although we cannot reject monotonicity statistically Perhaps the most interesting feature of these results is the decisive evidence that the cost functions for each sector differ substantially We must reject the null hypothesis that the technologies are identical between the two sectors imposed by constraining $\alpha_{11rs} = \alpha_{22rs}$ for all r, s) [28] This should not be oo surprising, since such a restriction is actually a special case of the eparable model in which the trade-off between outputs is linear, and we ave already rejected the more general form of separability

---

[26] In the base period 1958, when all commodity and factor prices are unity, factor shares are conveniently expressed as the sum of appropriate parameters as follows $\theta_{Ki} = \alpha_{iiKK} + \alpha_{iiKL} + \alpha_{iiKM}, \theta_{Li} = \alpha_{iiLL} + \alpha_{iiKL} + \alpha_{iiLM}, \theta_{Mi} = \alpha_{iiMM} + \alpha_{iiKM} + \alpha_{iiLM}$ Using the point estimates from table 1, column 2, we find that $\theta_{L1} = 77$, $\hat{\theta}_{K1} = 47$, $\hat{\theta}_{L2} = 43$, $\hat{\theta}_{K2} = 30$, from which $|\hat{\theta}| = +\ 0289$ Despite the nonlinearity of this statistic in the parameters, we can compute the approximate standard error from the formula $E(\theta - \hat{\theta})^2 = J' \sum J$, where $\sum$ is the variance-covariance matrix of estimated coefficients, and J is the vector of partial derivatives of θ with respect to the parameters evaluated at their estimated values (see Goldberger 1964) We obtain a calculated standard error of 00873 from which we conclude that θ is significantly greater than zero It is also easy to verify bv a similar procedure that $\hat{\theta}_{K1}\hat{\theta}_{M2} - \hat{\theta}_{K2}\hat{\theta}_{M1}$ is significantly greater than zero

[27] Only one pair of inputs can be complements when there are only three inputs Also, sign σ^{i}_{rs} = sign α_{iirs}

Special Restrictions

The negative and significant estimate for α_{11KL} is particularly disturbing because it implies the rather implausible possibility that capital and labor are complementary inputs for the nontraded-goods sector [29] This may be the result of aggregating over two types of output (services and structures) which have substantially different primary-factor intensities Suppose instead that we interpret this result as an indication of very limited flexibility in the choice of technique for this sector and impose the restrictions that all partial elasticities of substitution are zero in the production of nontradables This means constraining the parameter estimates for α_{11KL}, α_{11KM}, α_{11LM} to be zero and gives rise to the estimates presented in column 1 of table 3 The restrictions on input substitutibility make all but one of the remaining free parameters highly significant and the technology for the tradable-goods sector completely well behaved It also yields a marginal improvement in the Durbin-Watson statistic In addition, the calculated F-statistic with 3 and 76 df is 2 42, which is below the 5-percent critical value of 2 73. I conclude that we cannot reject the ull hypothesis of zero substitutibility in the nontraded-goods sector One nteresting feature of this model is the indication that, for the tradable-oods sector, capital and labor are close substitutes while capital and mports are complements These properties resemble those obtained for he one-sector (separable) model, and this is not surprising since our estrictions force all flexibility in the choice of inputs upon the one ector The final point to observe about the estimates for this case is ritical for purposes of making inferences about the income distributional ffects of higher tariffs Using the point estimates, we find that $\hat{\theta} = {}_{L1}\hat{\theta}_{K2} - \hat{\theta}_{K1}\hat{\theta}_{L2} < 0$, implying that services and structures are capital ntensive This reverses our previous results and suggests that higher ariffs raise real rentals and lower real wages [30]

One of the problems with both the general nonjoint model and the

---

[28] Identical cost functions for each type of output implies only six free parameters The esidual sum of squares under this restriction is 39 64, and the calculated F-statistic with and 76 df is approximately 88 0, implying decisive rejection of this hypothesis.

[29] For the nonjoint model, $\sigma^{i}_{rs} = C^{i}C^{i}_{rs}/C^{i}_{r}C^{i}_{s}$ where C^{i}, C^{i}_{r}, C^{i}_{s} are all positive for a ell-behaved technology Therefore, sign σ^{i}_{rs} = sign α_{iirs}

[30] Using the estimates presented in column 1, table 3, we find $\hat{\theta}_{L1} = 72$, $\hat{\theta}_{K1} = 52$, ${}_{L2} = 47$, $\hat{\theta}_{K2} = 26$ from which we calculate $|\hat{\theta}| = -0572$ Using the procedure utlined in n 26 above, the approximate standard error for this statistic is 0022, from hich we conclude that $|\theta|$ is significantly less than zero This, coupled with similar vidence that $\hat{\theta}_{K1}\hat{\theta}_{M2} - \hat{\theta}_{K2}\hat{\theta}_{M1}$ is significantly greater than zero implies strongly that igher tariffs make owners of capital services better off

TABLE 3

IZEF ESTIMATES OF GLGL JOINT COST FUNCTION UNDER ALTERNATIVE SUBSTITUTION RESTRICTIONS

	Fixed Coefficients for Services and Structures	Imports Only in Durables and Nondurables	Imports Separable from Primary Factors	Fixed Coefficients
α_{11KK}	51934 (02894)	1 13837 (8262)	$= \frac{\alpha_{11KL}\ \alpha_{11KM}}{\alpha_{11LM}}$	.50328 (0508)
α_{11LL}	72494 (02521)	1 60994 (9186)	$= \frac{\alpha_{11KL}\ \alpha_{11LM}}{\alpha_{11KM}}$	56968 (0468)
α_{11MM}	− 23384 (01975)		1 2450 (2507)	− 04848 (0374)
α_{11KL}		−0 85617 (8718)	0 65819 (0601)	
α_{11KM}			−0 74752 (0958)	
α_{11LM}			−0 69482 (1468)	
α_{22KK}	− 02742 (04705)	−0 76540 (6378)	$= \frac{\alpha_{22KL}\ \alpha_{22KM}}{\alpha_{22LM}}$	28973 (0407)
α_{22LL}	− 45027 (05671)	−1 05505 (7117)	$= \frac{\alpha_{22KL}\ \alpha_{22LM}}{\alpha_{22KM}}$	59084 (0403)
α_{22MM}	21878 (02684)	−0 05563 (0236)	−0 89010 (1865)	09613 (0298)
α_{22KL}	59229 (04342)	1 30594 (6717)	−0 06122 (0364)	
α_{22KM}	− 29706 (02918)	−0 07241 (0370)	0 26653 (0775)	
α_{22LM}	33001 (02540)	0 18882 (0371)	0 84830 (1057)	
D-W	1 5709	0 8182	0 9851	0 2505
SSR	5 5055	15 0938	10 0327	39 6424
SF	(26399)	(43711)	(35413)	(69530)
OBS	88	88	88	88

NOTE Standard errors in parentheses SSR = sum of squared residuals, OBS = observations

special zero-substitution case above is the implausible implication that the import requirement for the nontraded-goods sector is negative In the latter case, this aspect is particularly disturbing because the negative point estimate is highly significant The operational consequence of this result is that the least-cost input vector for a dollar of final output of services and structures at the given input prices actually consists of zero imported inputs and presumably more of both capital and labor services One might argue that our model should therefore explicitly incorporate the restriction that no factor input requirement can be negative It is also worth pointing out that the national-income accountant's treatment of all merchandise imports as final goods implies that no imports are required for the production of services and structures

To test the restriction that imports enter only as inputs in the production of durables and nondurables, we set $\alpha_{11KM} = \alpha_{11LM} = \alpha_{11MM} = 0$, and this gives rise to the estimates presented in column 2 The calculated F-statistic corresponding to the null hypothesis is distributed with 3 and 76 df and takes the value of 50 7 This is substantially above the 1-percent critical value of 4 06 and implies rejection of the null hypothesis that imports enter only as inputs for the production of tradable goods (durables and nondurables) Nevertheless, it is interesting to note that this restriction has the force of removing the negative and significant estimate for the elasticity of substitution between capital and labor in the nontradable-goods sector (services and structures) The point estimate for α_{11KL} remains negative, but it is insignificant from zero It is also worth pointing out that the indication of substantial flexibility in the choice of technique for the production of tradables remains true for this model Capital services and imports appear to be complementary inputs, while capital and labor are highly substitutable Finally, nontradable goods appear from this model to be labor intensive, as was the case for the general nonjoint model Higher tariffs would then appear to distribute income from capital to labor, if this model were valid

I have examined two technological restrictions on the nonjoint model in an effort to gain further insight into the specific character of the technology for each sector This procedure is motivated by the failure of the unconstrained nonjoint model to exhibit all the characteristics of a well-behaved technology While the restrictions I have imposed have no strong a priori claim to empirical validity, it is useful to examine their implications for other characteristics of the model, we can also obtain a reasonable indication of the empirical validity of the restrictions if we believe the classical theory of statistical inference I now proceed to examine two further restrictions on the nonjoint model before concluding the paper

When the technology is nonjoint, it is useful to examine whether imported materials are functionally separable from primary factors (capital and labor) in each productive process This property has been an implicit maintained hypothesis of previous efforts to estimate economy-wide value-added production processes Without this separability property, a consistent index of value added does not exist for each sector Separability between primary factors and imported materials implies equality between the partial elasticities of substitution between imports and capital services on the one hand and imports and labor services on the other[31] In terms of the parameters of the nonjoint model, the following two sets of mildly nonlinear restrictions obtain $\alpha_{iiKK}/\alpha_{iiKL} = \alpha_{iiKL}/\alpha_{iiLL} = \alpha_{iiKM}/\alpha_{iiLM}$ $(i = 1, 2)$[32] Column 3 of table 3 presents the

---

[31] See Berndt and Christensen (1973)

[32] Separability between primary factors and imported materials in the two-sector nonjoint model implies the conditions $C_r^iC_{wp}^i = C_w^iC_{rp}^i$ $(i = 1, 2)$ For the GLGL form we have $C_r^iC_{wp}^i = \alpha_{iiKK}\alpha_{iiLM}(w\ p)^{-1/2} + \alpha_{iiKL}\alpha_{iiLM}(r\ p)^{-1/2} + \alpha_{iiKM}\alpha_{iiLM}(r\ w)^{-1/2} = C_w^iC_{rp}^i = \alpha_{iiLL}\alpha_{iiKM}(r\ p)^{-1/2} + \alpha_{iiKL}\alpha_{iiKM}(w\ p)^{-1/2} + \alpha_{iiLM}\alpha_{iiKM}(r\ w)^{-1/2}$ $(i = 1, 2)$ Equating parameters for identical terms gives the conditions in the text

accession numbers HU

estimates of the eight free parameters corresponding to the nonjoint model with imports separable from capital and labor in both sectors. The calculated F-statistic with 4 and 76 df is 19.0, which is above the 1-percent critical value of 3.58, implying rejection of the null hypothesis of functional separability of imported materials for both sectors. Despite the evidence against this restriction, it is interesting to note that it actually reverses the degree of capital-labor substitutibility between the two sectors. Neither sector exhibits the characteristics of a well-behaved technology with this restriction imposed, and the income distributional effects of higher tariffs are reversed, the return to capital increases with higher tariffs since services and structures are capital intensive.

The generalized Leontief form becomes a fixed coefficients technology when all off-diagonal parameters are zero. To test the hypothesis of zero input substitution in both sectors, we set $\alpha_{iirs} = 0$ whenever $r \neq s$, and this gives rise to the estimates presented in column 4 of table 3. The calculated F-statistic corresponding to the null hypothesis is distributed with 6 and 76 df and takes the value of 87.6. Since this is substantially above the 1-percent critical value of 3.07, we must decisively reject the null hypothesis of a nonjoint, fixed-coefficients technology. It is worth observing that the estimates corresponding to this model also imply that services and structures are relatively capital intensive while using relatively little imported materials. The inference from this model again is that higher tariffs distribute real income from labor to capital.

V. Conclusion

The results of this paper regarding the income distributional effects of tariffs are somewhat inconclusive. Both special forms of the technology which have frequently been adopted as maintained hypotheses in previous theoretical and empirical work are shown to be of questionable validity for purposes of aggregative analysis. If, however, either form is adopted a priori, then they yield similar implications regarding the factoral protection afforded by the U.S. tariff. Both the separable and the unrestricted nonjoint models imply that higher tariffs distribute real income from capital to labor for the postwar U.S. economy. The reasoning behind this result differs for each model. For the separable technology, it is purely a consequence of the evidence of complementarity between capital services and imports (higher tariffs raise import prices and stimulate the demand for labor services while reducing the demand for capital services). This suggests that the goods which the United States imports require relatively large amounts of labor input and relatively little capital input if produced locally. For the nonjoint technology, the sector that is labor intensive (services and structures) also uses relatively little imported materials. This means that higher tariffs tend initially to raise unit costs more in the capital-intensive sector (durables and nondurables). Since output prices are fixed by the assumption of price-taking behavior, the ultimate consequence of this is that rental rates must fall and wage rates must rise to restore equilibrium.

The inferences obtained from both the separable and nonjoint models must be regarded with some skepticism, because if we proceed from the maintained hypothesis of the general GLGL joint cost function, both forms are rejected statistically at the 5-percent level The evidence against nonjointness is not nearly as strong as that against separability. However, we find that the inferences about factoral protection are very sensitive to special restrictions imposed on the nonjoint model in an effort to isolate the specific characteristics of the technology for each sector The most dramatic example of this occurs when we adopt the assumption of zero-input substitutibility for the services and structures sector This restriction cannot be rejected statistically, and yet it actually reverses the factor-intensity condition, suggesting that services and structures are capital intensive The consequence of this is that higher tariffs distribute income from labor to capital rather than vice versa! The sensitivity of estimated factor intensities to prior restrictions is not too surprising since it is an open empirical question whether services and structures treated together are on balance relatively labor intensive vis-à-vis the rest of the economy We may be guilty of a serious misspecification error in lumping two categories of output whose factor intensities differ dramatically

Finally, we add the caveat that our inferences are as valid as the assumption of exogenous output prices If one of our outputs is indeed a nontraded good, then its price is endogenously determined in a complete general equilibrium model which includes the consuming sector It is interesting to consider how the endogeneity of the price of nontradables would alter our inferences First, nothing changes within the context of the separable model, since any changes in output prices have no impact on factor prices For the nonjoint model, the income distributional effects will be either reinforced or partly offset by any subsequent market-clearing change in the price of nontradables The effect of higher import prices on the output mix at given output prices depends on elasticities of substitution as well as factor intensities, and our empirical estimates are quite ambivalent about this matter However, it can be shown that if all inputs are substitutes in both sectors, and if higher tariffs raise wages and lower rentals, then the output of nontraded goods will expand if it is labor intensive[33] Thus, higher tariffs will create an initial excess supply of nontraded goods, a subsequent reduction in the relative price of nontradables, and a partially offsetting reduction in wage rates. The secondary changes in factor prices are unclear whenever there is some complementarity in the use of factors in either sector Because we have some evidence of this, I prefer to conclude the paper on the rather agnostic note that precise knowledge of factor intensities and elasticities of substitution is needed in order to draw complete and accurate inferences about income distributional effects when one good is not traded Nevertheless, the "first-order" effects of higher tariffs measured in this paper are likely to be accurate the more the producing sector approximates a price-taking firm

---

[33] The effects of higher import prices on outputs can be determined from total differentiation of the full-employment conditions $y_1C_w^1 + y_2C_w^2 = L$, and $y_1C_r^1 + y_2C_r^2 = K$ Letting $dw/dp = \delta_w$ and $dr/dp = \delta_r$, the following expressions are obtained: $dy_1/dp = (V_{rr}\delta_r + V_{rw}\delta_w + V_{rp})/\Delta$, and $dy_2/dp = (V_{rw}\delta_r + V_{ww}\delta_w + V_{wp})/\Delta$ where $\Delta \equiv C_w^1C_r^2 - C_r^1C_w^2 > 0$ if the first sector is labor intensive In these expressions, $V_{rw} = y_1C_{rw}^1 + y_2C_{rw}^2 > 0$ when capital and labor are substitutes. Assuming $\delta_w > 0$ and $\delta_r < 0$, it is easy to verify that $dy_1/dp > 0$ whenever all inputs are substitutes.

References

Allen, R D G *Mathematical Analysis for Economists* London Macmillan, 1938

Arrow, K J , Chenery, H B , Minhas, B S , and Solow, R M "Capital Labor Substitution and Economic Efficiency " *Rev. Econ and Statis* 43, no 3 (August 1961) 225-50

Balassa, B "Tariff Protection in Industrial Countries An Evaluation " *J P E* 73, no 6 (December 1965) 573-94

Baldwin, R E "Determinants of the Commodity Structure of U S Trade " *A E R* 61, no 1 (March 1971) 126-46

Basevi, G "The U S Tariff Structure An Estimate of Effective Rates of Protection of U S Industries and Industrial Labor " *Rev Econ and Statis* 48, no 2 (May 1966) 147-60

Berndt, E R , and Christensen, L R "The Internal Structure of Functional Relationships Separability, Substitution, and Aggregation." *Rev Econ Studies* 40, no 3 (July 1973) 403-10

Burgess, D F "Production Theory and the Derived Demand for Imports." *J Internat Econ* 4, no 2 (May 1974) 103-18 (*a*)

——— "A Cost Minimization Approach to Import Demand Equations " *Rev Econ and Statis* 56, no 2 (May 1974) 225-34 (*b*)

Chow, G , and Fair, R "Maximum Likelihood Estimation of Linear Equation Systems " Research Memorandum no 118, Princeton Univ , 1971

Christensen, L R , and Jorgenson, D W "U S Real Product and Real Factor Input 1929-1967 " *Rev Income and Wealth* 16, no 1 (March 1970) 19-50

Diewert, W E "An Application of the Shephard Duality Theorem A Generalized Leontief Production Function " *J P E* 79, no 3 (May/June 1971) 481-507

——— "Applications of Duality Theory " In *Frontiers of Quantitative Economics*, edited by R D Intriligator and D A Kendrick Vol 2 Amsterdam North-Holland, 1974

Goldberger, A S *Econometric Theory* New York Wiley, 1964

Hall, R E "The Specification of Technologies with Several Kinds of Outputs " *J P E* 81, no 4 (July/August 1973) 878-92

Johnson, H G *Two-Sector Model of General Equilibrium* Chicago Aldine, 1971

Jones, R W "Effective Protection and Substitution " *J Internat Econ.* 1, no. 1 (February 1971) 59-81

Jorgenson, D W , and Griliches, Z "The Explanation of Productivity Change " *Rev Econ Studies* 34, no. 3 (July 1967). 249-83

Khang, C "Factor Substitution in the Theory of Effective Protection: A General Equilibrium Analysis." *J Internat. Econ.* 3, no 3 (August 1973). 227-44

Kmenta, J., and Gilbert, R. "Small Sample Properties of Alternative Estimators of Seemingly Unrelated Regressions." *J. American Statis. Assoc.* 63 (December 1968) 1180–1200.

McFadden, D L "Cost Revenue and Profit Function " In *An Econometric Approach to Production Theory*, edited by D L. McFadden. Amsterdam· North-Holland, 1975

Mayer, W "Short-Run and Long-Run Equilibrium for a Small Open Economy " *J P E* 82, no 5 (September/October 1974). 455–68.

Melvin, J R "Production and Trade with Two Factors and Three Goods " *A E R* 58, no 5 (December 1968) 1249–68.

Samuelson, P. A "Prices of Factors and Goods in General Equilibrium " *Rev Econ Studies* 21, no 1 (January 1954) 1–20

Shephard, R W *Cost and Production Functions* Princeton, N J Princeton Univ Press, 1953

Stolper, W , and Samuelson, P A "Protection and Real Wages " *Rev Econ. Studies* 9, no 1 (November 1941) 58–73

Travis, W P *The Theory of Trade and Protection* Cambridge, Mass Harvard Univ Press, 1964

——— "The Effective Rate of Protection and the Question of Labor Protection in the United States " *J P E* 76, no 3 (May/June 1968) 443–61

——— "Production, Trade, and Protection When There Are Many Commodities and Two Factors " *A.E R* 62, no 2 (March 1972) 87–106

U S Department of Commerce, Office of Business Economics *The National Income and Products Accounts 1929–65.* Washington Government Printing Office, 1966

Uzawa, H "Production Functions with Constant Elasticities of Substitution " *Rev Econ Studies* 29, no 3 (July 1962) 291–99

Woodland, A D "Duality Principles in International Trade " Discussion Paper no 73–22, Univ British Columbia, December 1973

Zellner, A "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias " *J American Statis Assoc* 57, no 2 (June 1962) 979–92

, Experience, Earnings, and stments in Human Capital

nders Klevmarken
versity of Stockholm

ohn M. Quigley
ale University

This paper explores the distinction between physical age and experience within the context of the human capital model of earnings determination. We extend the human capital model to distinguish the investment behavior of age-experience cohorts and to incorporate the effects of calendar time. Several alternative models are developed which are empirically tested based on a unique body of data (on Swedish engineering graduates) reflecting the earnings of cohorts with identical educational qualifications cross-classified by age and experience for a 10-year period. The results clearly indicate the necessity for distinguishing between experience and age in analyses of investment in human capital and are consistent with the interpretation that younger members of the same experience cohort are more efficient in producing human capital The analysis also indicates the sensitivity of investment profiles to a priori restrictions on functional form and detects an inflection point in earnings increases consistent with Becker's original formulation of human investment theory

Most studies of the human capital investment process relate cross-sectional earnings profiles to schooling levels and years of experience, usually measured by calendar age. In this paper we analyze the conceptual distinction between calendar age and years of experience in explaining investment behavior and in interpreting earnings profiles.

This distinction was first noted by Mincer (1970) in his classic review article, it is also an integral part of his most recent work (Mincer 1974), where he shows that identical experience profiles of investment can lead

---

Facilities for this research were provided by the Industrial Institute for Economic and Social Research, Stockholm, and the National Bureau of Economic Research. This paper is not, however, an official document of either organization. We are grateful to Gary Fields and Eric Hanushek for useful discussions

[*Journal of Political Economy*, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

to substantial differences in age-earnings profiles if individuals receive unequal amounts of schooling

Our analysis goes one step further in considering the investment decisions of individuals who receive identical amounts of schooling but complete their schooling at different ages Variations in the age at which schooling is completed are substantial, at least for graduate education For example, Harris (1972, p 89) reports that the mean time elapsed between baccalaureate and doctorate degrees in the United States is about 10 years, with a standard deviation of 6 years. Harris also presents separate estimates of the mean and standard deviation of time elapsed between baccalaureate and doctorate degrees for 13 fields of study The coefficients of variation range between 45 (for physics) and 56 (for mathematics) There is also evidence of considerable variation in the age at which bachelor degrees are awarded [1]

In this paper we indicate, by extending Mincer's analysis of the human capital model, how investment behavior and capital depreciation depend on both age and years of experience even for individuals with identical quantities of education A general model is developed (Sec I), distinguishing between cohort and cross-sectional earnings profiles, which relates observed earnings to investment and disinvestment in human capital, to returns from the human capital stock, and to those secular increases in productivity which are unrelated to investment behavior

By making alternative assumptions about these factors, we derive two specific models for subsequent testing (Sec II) The models differ in generality, making possible an assessment of the realism of several underlying assumptions

In Section III the models are estimated separately for two populations using annual cross sections of earnings classified by age, years of experience, and calendar year For several reasons described below, the data (obtained from annual surveys of all engineering graduates in Sweden) are particularly well suited to testing the theoretical model

In the analysis of Sections IV and V, the estimated earnings profiles are first presented, followed by an analysis of the implied human capital investment behavior The framework also allows an investigation of the differentiation of starting salaries for cohorts which enter the full-time labor market at different ages

I. The Economic Model

According to human capital theory, individuals make conscious or unconscious decisions to invest in their own productive capacity Individuals invest first in length and type of schooling. After entering the labor market, they make additional investments in training, health care, and possibly additional formal schooling These human capital investments and their rates of return are important determinants of individuals' earnings profiles

---

[1] E g , a sample of U S graduate students analyzed by Davis revealed that more than one-third were *not* between the ages of 21 and 23 when they completed their undergraduate education The mean age was 23 years, with a standard deviation of 3 years, and 4 percent of the sample received bachelor degrees at age 20 or younger (Davis 1962, p 183)

In its simplest form, the human capital model postulates that

$$E_j = E_{j-1} + r_{j-1}C_{j-1}, \tag{1}$$

where E_j is potential earnings of an individual with j years of experience, that is, the earnings obtainable if no resources are used for investment in human capital in the jth year, C_{j-1} is the amount by which potential earnings are reduced by human capital investment in the $j - 1$th year of experience, and r_{j-1} is its corresponding marginal rate of return

The postulate of diminishing returns to the production of human capital in any period implies that gross investment in human capital will be spread out over the working life In addition, even if the marginal cost-of-investment curve facing an individual were identical during each period of possible investment activity, the finiteness of working life would ensure that the marginal revenues of later investments were lower than those of earlier investments Taken together, these considerations imply that the profile of human capital investments throughout the working life will generally be declining

Several factors complicate this simple relationship between earnings and human capital investment, including (1) the depreciation of human capital, (2) the complex dependence of human capital investment on "ability" and physical age as well as years of experience, and (3) the presence of physical capital

1 As an analogue to physical capital, human capital is subject to deterioration The existence of retirement alone points to this conclusion, but the precise level of age-related depreciation probably varies with an individual's occupation or industry For example, it is widely observed that piece-rate output declines for older workers, and we may expect mental abilities to decline at some point in analogous fashion Knowledge and skills may also become obsolete Thus, depreciation may be related to the vintage of the human capital stock

2 In the "naive" formulation of equation (1), physical age and experience are used interchangeably But even if we consider a population with given educational qualifications, the "naive" formulation would only be appropriate (*a*) if all individuals were continuously employed in the labor force after completing schooling and (*b*) if all individuals completed schooling and entered the labor force at the same age

The well-known problems of applying the human capital model to females illustrate the first qualification For females, the cloudy relationship between physical age and years of experience, coupled with our

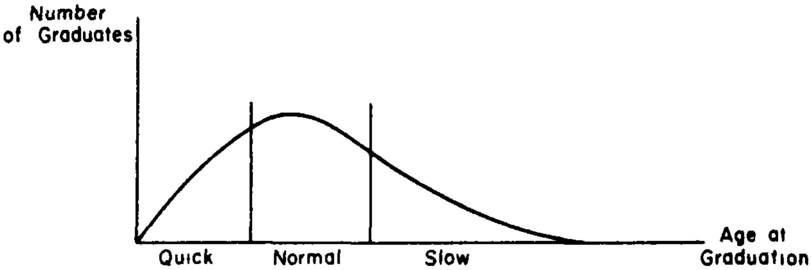


FIG 1 —Hypothetical frequency distribution of graduation age for a cohort of workers

ignorance of human capital investment behavior during periods of non-participation in the labor market, makes the "naive" model inapplicable [2]

However, even if all individuals were continuously employed after completing schooling, variations in the age at which schooling is completed will still complicate the analysis

For postsecondary education, real variations exist in the age at which training is completed and individuals enter the full-time labor market Variation in graduation age is most pronounced for holders of advanced degrees, but there also exists variation in the age of completion of most educational qualifications Individuals may "drop out" for long or short periods on the way to their degrees, their training programs may be interrupted by military service, or circumstances may necessitate their taking part-time jobs on the way to completing their educational qualifications

Figure 1 presents a hypothetical frequency distribution of age at graduation for a cohort of workers If we stratify the distribution in figure 1 into, say, three groups—a "quick" group, composed of those who are younger than "normal" at graduation, a "normal" group, who graduate at the "normal" age, and a "slow" group, composed of those who are older than "normal" at graduation—we may roughly characterize the individuals in each group in the following way The "quick" group is made up of students who are more efficient in learning than average but who graduate with no previous labor market experience, the "normal" group contains students of average ability who graduate with no (or very little) previous labor market experience, and the "slow" group contains those who are perhaps less efficient in learning than average but are often those with additional labor market experience In general, one would expect that students who are efficient in learning are also efficient in productive work For this reason their *potential* starting salaries would tend to be high However, previous labor market experience would also increase their potential starting salaries

---

[2] See, e g , Mincer and Polachek (1974) for an application of the theory to female earnings

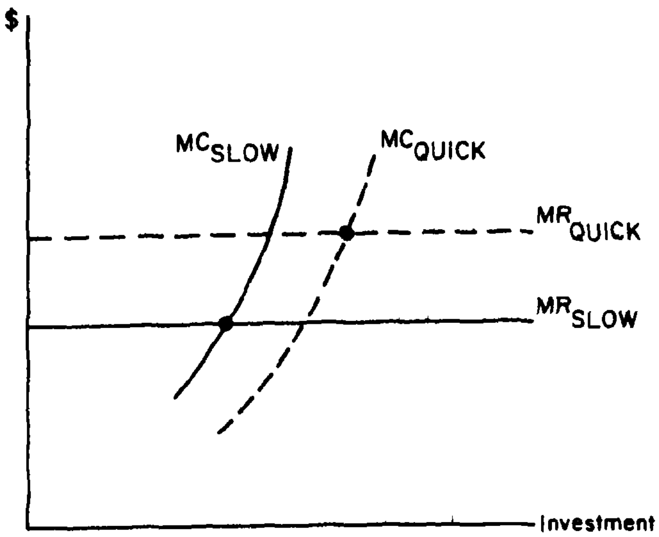


FIG 2 — Marginal costs and revenues for younger and older workers with the same xperience

The quick group should thus have a relatively high *potential* starting alary But it is more difficult to say anything about their *observable* tarting salaries, since, with greater efficiency in learning, they should nvest more in human capital than the others The slow group may also eceive relatively high potential and actual starting salaries as a result f their previous labor market experience

Figure 2 illustrates how physical age also influences postschool nvestment behavior The diagram indicates that depreciation of human pital and variations in age and experience can affect analyses of human pital investment profiles by altering the marginal cost-of-investment rve and/or the marginal revenue curve facing different cohorts of dividuals

At the same level of experience, there are at least three factors which ay cause the marginal cost curve facing an older worker to lie above the rve facing a younger worker First, if the age of an individual is an put in the production function for human capital and its marginal roduct is negative, then *ceteris paribus* the cost of human capital will be eater for older workers Second, if age at graduation measures efficiency learning, then the marginal cost curve for older workers will lie above e curve for younger workers Third, if the production function for uman capital includes the current capital stock as an input (see Ben-orath 1967, 1970), then any age-related depreciation will reduce the urrent stock more for older than for younger workers

The effect on the marginal revenue curve is more straightforward At a given level of experience, the marginal revenue curve is lower for older workers If all individuals retire from employment at a common age, the marginal return from any investment is lower for older workers However, even if the age of retirement is variable, the marginal revenue curve for older workers will be below that for younger workers if depreciation is an increasing function of age [3] In addition, if those with more "ability," as measured by age at which schooling is completed, have lower depreciation rates, this is another reason for expecting lower marginal revenue curves for older members of an experience cohort

All of these factors suggest that investment ratios decline with experience and with age But this tendency for declining investment ratios need not logically exist for all intervals in the range of age and experience for all individuals For example, if there are intervals where productivity in learning grows faster than productivity in earning, investment ratios may increase Presumably, these intervals occur during the early years in the labor market, if at all (see Mincer 1974, pp 13–16) [4]

3 For individuals who enter the labor market in different calendar years, we may expect starting salaries to vary, causing shifts in earnings profiles In the short run, these variations in starting salaries are responsive to supply and demand conditions in particular labor markets In the long run, however, the systematic increases in starting salaries accruing to those entering the labor force later but with the same embodied human capital reflect economic growth, that is, increases in physical capital, its productivity and organization, and other factors distinct from human capital investments Similarly, the increases in earnings received after entering the labor force depend on net investments in physical capital as well as the returns to human capital investment

We now extend the human capital model to confront these issues First, we develop a formal framework to incorporate depreciation of human capital, variations in experience and physical age, and variations in initial salaries By making a number of specific assumptions, we derive two estimable models which are investigated below in several variants

The general model is developed by describing several definitional relationships among *potential starting salaries*, *potential earnings*, and the actual *observable earnings* received by individuals in the labor market

Define the *potential starting salary* (i e , the starting salary obtainable if no resources are expended in human capital investment) for individuals who entered the labor force in year b at the age of k_0 as

---

[3] Only in the special case where depreciation is at a constant rate and retirement age is variable would the marginal revenue curves coincide

[4] Since the theory implies that there is a secular but not necessarily monotonic decline in investment ratios with age and experience, empirical analyses which impose some "convenient" functional form (e g , linear in experience) throughout the age or experience range are unlikely to detect intervals where investment ratios are increasing, if they exist at all Indeed, the empirical results reported in Sec IV allow an investigation of this question

$$E_{bk_00} = E_{0k_00} \exp\left(\sum_{t=1}^{b} \beta_{tk_00}\right), \tag{2}$$

where E_{0k_00} is a function of k_0 and β_{tk_00} is simply the rate of increase of the starting salary between year $t - 1$ and t

Define the *potential earnings* (i e , the salary obtainable if no resources are expended in human capital investment) at time $b + j$ after j years of experience for those who entered the labor force at age k_0 as

$$E_{bk_0j} = E_{bk_0j-1}[1 + \beta_{bk_0j} + r_{j-1}(c_{bk_0j-1} - d_{bk_0j-1})], \tag{3}$$

where β_{bk_0j} is the rate of increase in potential earnings independent of investments in human capital, r_{j-1} is the rate of return to the previous year's investments, c_{bk_0j-1} is the *investment ratio*[5] (i e , the ratio between the previous year's investment in human capital and potential earnings for that year), and d_{bk_0j-1} is the corresponding depreciation ratio

Since by definition human capital investment in any year is the difference between potential earnings and observable income,

$$Y_{bk_0j} = E_{bk_0j}(1 - c_{bk_0j}), \tag{4}$$

the basic model can be stated simply as

$$\begin{aligned} Y_{bk_0j} = {} & E_{0k_00} \exp\left(\sum_{t=1}^{b} \beta_{tk_00}\right)(1 - c_{bk_0j}) \\ & \times \prod_{i=1}^{j} [1 + \beta_{bk_0i} + r_{i-1}(c_{bk_0i-1} - d_{bk_0i-1})] \end{aligned} \tag{5}$$

By taking the natural logarithm of both sides of equation (5), the following approximate expression is obtained

$$\begin{aligned} \log Y_{bk_0j} = {} & \log E_{0k_00} + \sum_{t=1}^{b} \beta_{tk_00} + \sum_{i=1}^{j} \beta_{bk_0i} \\ & + \sum_{i=1}^{j} r_{i-1}(c_{bk_0i-1} - d_{bk_0i-1}) - c_{bk_0j} \end{aligned} \tag{6}$$

Equation (6) expresses in a general form the relationship among the observed incomes obtained by individuals, their starting salaries, their human capital stocks, and the rates of return to their human capital investments

---

[5] The terminology and the formulation of investment and depreciation ratios are due to Becker and Chiswick (1966)

II. Alternative Specifications

To derive an estimable model from equation (6), we must impose specific assumptions about the functional relationships implied by β, c, d, and E. Assumptions 1–5 below yield two specific and plausible forms of the general model

Assumption 1 rates of return, r

$$r_j = r \qquad (7\ \text{I}),\ (7\ \text{II})$$

For both models we assume that the rate of return to human capital investment is the same for all years of experience, equivalently, there exists a market for human capital This assumption is commonly and almost uniformly adopted in empirical applications of the human capital approach

Assumption 2 rates of earnings increase unrelated to investment, β

$$\beta_{tk_0j} = \beta, \qquad \text{for all } t,\ k_0, \text{ and } j > 0, \qquad (8\ \text{I})$$

$$\beta_{tk_0j} = \beta(k_0), \qquad \text{for all } t \text{ and } j > 0 \qquad (8\ \text{II})$$

For model I we assume that the rate of increase in earnings independent of investment in human capital is constant Thus we abstract from changes in the growth rate of the economy cyclical changes in the rate of earnings increase or particular features of narrowly defined labor markets This simplification may seem unrealistic, but no bias will be introduced into estimates of cohort age-earnings profiles as long as β_{tk_0j} is independent of years of experience, j

For model II we allow the rate of increase to vary systematically with graduation age This extension is included because, as we noted above, graduation age may be an indication of "ability", more-able individuals are more efficient in utilizing new capital equipment, new methods of production, and so forth

Assumption 3 investment profiles, c

$$c_{bk_0j} = c_1(j) + c_2(k), \qquad k = k_0 + j, \qquad (9\ \text{I})$$

$$c_{bk_0j} = c(j, k) \qquad (9\ \text{II})$$

For model I we assume that the investment ratio can be expressed additively as two separable functions, one relating experience and one relating age to investment behavior This implies that the difference in gross investment ratio between two workers of the same age but at different levels of experience does not systematically depend on their age, and vice versa Figure 3 illustrates this property of model I in the case of linearly declining investment ratios The C_1C_1 in the figure represents the profile of investment ratios as a function of experience To this we add the profile of investment as a function of physical age The three dashed

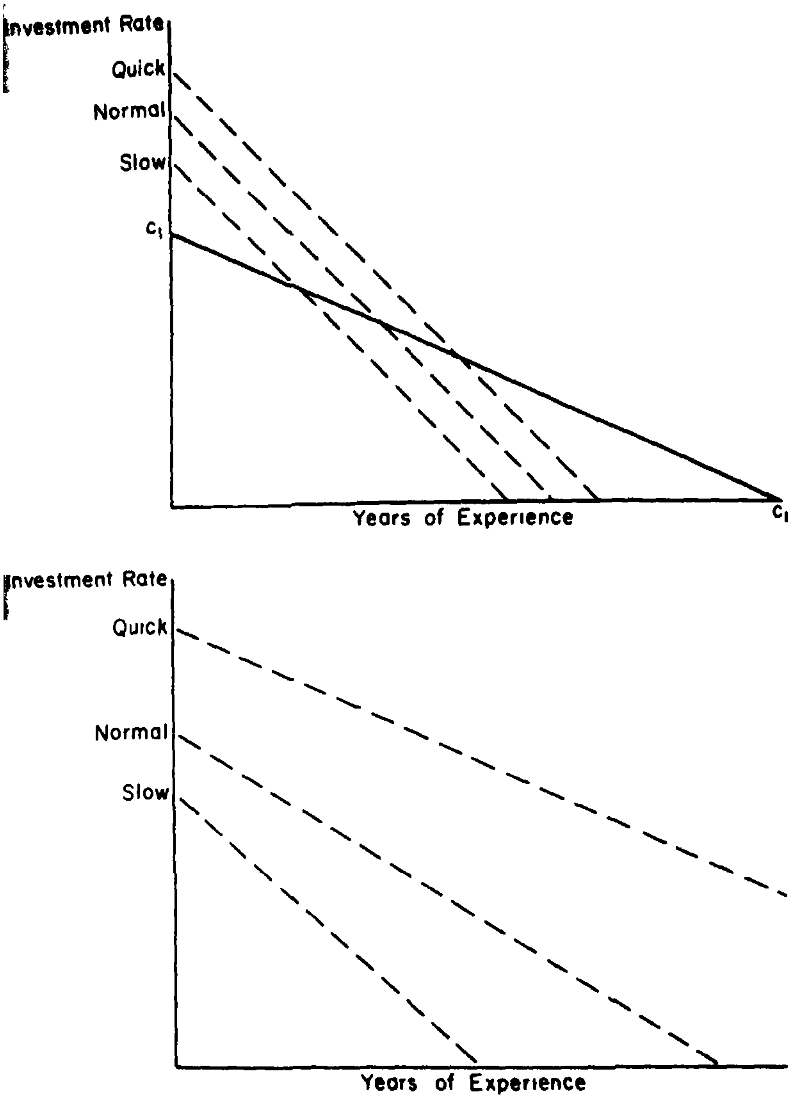


Fig 3 —Alternative models of capital investment

es represent the resulting profiles of investment ratios for three pothetical groups of individuals, those who graduate when they are unger than the normal age ("quick"), at the normal age ("normal"), d older than the normal age ("slow") The profiles are all parallel, the difference in graduation age between the quick and normal groups the same as between the normal and slow groups, the profiles are also equidistant If the investment ratios are nonlinear functions of experience and age, respectively, the profiles will no longer be parallel, but the difference between any two profiles will still be independent of experience

For model II, we allow for interactions between age and experience This is illustrated in figure 3 for the case of linear functions In this hypothetical figure, those who graduate relatively younger have a higher initial investment ratio than those who graduate at an older age, and their profiles have a smaller slope Our theory suggests that the first group invests more than the second This, however, does not necessarily imply that the investment rate is higher, since the potential starting salary may well be higher for those who graduate when they are younger

Assumption 4 depreciation profiles, d

$$d_{bk_0j} = d_1(j) + d_2(k), \tag{10 I}$$

$$d_{bk_0j} = d(j, k) \tag{10 II}$$

Human capital may depreciate in several ways, due to the obsolescence of skills or the deterioration of mental and physical capacity Both models recognize deterioration by relating annual depreciation to physical age In addition, since much of postschool investment is made during the first years in the labor market, years of experience reflect the obsolescence of capital of a given vintage, that is, it measures the elapsed time since schooling and most of postschool investments were made

We expect the profile of depreciation rates to be nondecreasing, the functional relationships for models I and II parallel those described for the investment profile

Assumption 5 potential starting salaries, E

$$\log E_{bk_00} = \alpha + \beta b + r \sum_{k=0}^{k_0-1} [c_2(k) - d_2(k)], \tag{11 I}$$

$$\log E_{bk_00} = E(k_0) + [\beta(k_0)]b \tag{11 II}$$

For model I the assumption implies that cross-sectional log earnings profiles will exhibit a parallel shift between any two years [6] Model I emphasizes the possibility of labor market activity (and human capital investment) prior to graduation, the specification assumes that those who are older when they graduate have invested more in human capital than those who are younger and thus receive higher starting salaries

For model II the rate of increase in starting salaries is a function of graduation age The assumption implies that potential and observable starting salaries may vary quite generally with graduation age, k_0

---

[6] In other words, the rate of increase in salary due to factors other than investments in human capital is the same for each year of experience and age and is the same as the rate of increase in starting salaries

The two models of human capital investment behavior imply observed earnings profiles:

$$\log Y_{bk_0j} = \alpha + \beta(b + j) + r\sum_{i=0}^{j-1}[c_1(i) - d_1(i)] + r\sum_{i=0}^{k_0+j-1}[c_2(i) - d_2(i)] - c_1(j) - c_2(k_0 + j), \qquad (12\ \mathrm{I})$$

$$\log Y_{bk_0j} = E(k_0) + [\beta(k_0)](b + j) + r\sum_{i=0}^{j-1}[c(i, k_0 + i) - d(i, k_0 + i)] - c(j, k_0 + j) \qquad (12\ \mathrm{II})$$

The shape of the earnings profiles implied by the two models is revealed by the following expressions for the rate of increase in earnings between the $j - 1$ and the jth year of experience for those starting to work in year b at k_0 years of age, say, γ_{bk_0j}

$$\gamma_{bk_0j} = \beta + r[c_1(j - 1) + c_2(k_0 + j - 1) - d_1(j - 1) - d_2(k_0 + j - 1)] - [c_1(j) + c_2(k_0 + j) - c_1(j - 1) - c_2(k_0 + j - 1)], \qquad (13\ \mathrm{I})$$

$$\gamma_{bk_0j} = \beta(k_0) + r[c(j - 1, k_0 + j - 1) - d(j - 1, k_0 + j - 1)] - [c(j, k_0 + j) - c(j - 1, k_0 + j - 1)] \qquad (13\ \mathrm{II})$$

As noted above, increasing investment ratios are conceivable for some intervals in an individual's career, but they must ultimately decline with age or experience, depreciation rates may increase. Equations (13) indicate the implications for earnings profiles more generally, an increase in earnings will be observed as long as the investment ratio does not increase more than the return to the previous year's investment.

Consideration of equations (13) leads to the following conclusions (i) If investment ratios decline at a constant rate or at a decreasing rate, the rate of increase in earnings *decreases*. (ii) If investment ratios decline at a rapidly increasing rate of change (greater than the rate of return), the rate of increase in earnings *increases*. (iii) If investment ratios decline at a moderately increasing rate of change (less than the rate of return), the rate of increase in earnings *decreases*. For example, in model I, if investment and depreciation ratios are all linear functions, it follows that those who graduate at a younger age will always have larger increases in earnings than those who graduate at an older age. However, if c_2 is not a linear function, this is not necessarily true.[7] In contrast to this

---

[7] E.g., if the investment ratio is large for young people but levels off rapidly with physical age (a step function is the limiting case), it is possible that those who graduate at younger ages will experience a smaller increase in earnings during their first years in the labor force than those who graduate at older ages. After some years of experience, however, the reverse should be true.

The data to be used are geometric average salaries, thus, the model should be stated in average form. Since we assume that individual earnings are independent, this is easily accomplished

$$\operatorname{var}(\bar{Z}_{bk_0j}) = \frac{(1+j)\sigma^2}{n_{bk_0j}}, \tag{18}$$

where $\bar{Z}_{bk_0j}$ represents the mean logarithmic error and n_{bk_0j} is the number of individuals comprising the geometric averages.

For all specifications except those involving profile B, the models are estimated by generalized least squares. However, when model I is estimated using profile B, the exponential investment function (eq. [14 B]), the equation contains a nonlinear parameter. Equation (14 B) has therefore been estimated by the golden-section search method (Flanagan 1969), that is, by minimizing the residual sum of squares over a range of values for the nonlinear parameter V.

III. Results

The two models are tested empirically using salary data obtained from the Swedish Association of Graduate Engineers. These data include geometric average salaries by age, year of graduation, and calendar year for separate populations of electrical and mechanical engineers. The data cover the 10-year period 1961–70 and are obtained from an annual survey of all engineering graduates in Sweden.[10] The data thus refer to those with identical educational qualifications belonging to well-defined professions, this makes the assumption of identical rates of return to human capital investments slightly less heroic than in other studies. Moreover, the data measure wages and salaries (not total income) in a society where, for all practical purposes, individual variations in health-care investment can be neglected.

The dependent variable is the logarithm of average monthly salaries in constant (1949) Swedish kronor. We assume that the year of entering the labor force (k_0) is the same as the year of graduation, that is, years of experience equal the number of years elapsed since graduation.[11] For profile C, there are nine experience variables for electrical engineers, reflecting the number of years in each interval. In particular there is one interval for each of the first 4 years of experience. For mechanical engineers, the data allowed 10 experience variables to be defined. Again, there is one interval for each of the first 4 years of experience. Seven physical age intervals are included for electrical engineers, and eight are included for mechanical engineers.

---

[10] The response rate for these surveys varied between 76 and 94 percent during the period (see Klevmarken [1972, chap. 2] for a discussion of the underlying data). A total of 1,130 observations on (geometric) average salaries is available for electrical engineers, the sample size is 1,042 for mechanical engineers. In 1965, e.g., the averages were obtained from surveys of 1,462 electrical engineers and 1,416 mechanical engineers.

[11] The underlying populations during this period were approximately 99 percent male, and the unemployment rate for engineers was about 0.5 percent.

For model I, the coefficient estimates are presented in tables 1 and 2. For each profile A, B, and C, two sets of results are presented, one including the estimated coefficients of the physical age variables and one based simply on the experience variables.

For each of the three profiles, the magnitudes of the coefficients of the experience variables change when physical age is excluded, implying a less steep earnings profile in the experience dimension. The physical age variables explain an additional 5 percent of the variance for both samples.

The general pattern of the estimated coefficients is the same for both samples, but a minor difference in the pattern of the physical age coefficients occurs in the age interval 55–59 years for mechanical engineers, where the estimates are larger than expected.

For electrical engineers, the real increase in starting salaries is estimated to be about 1.5 percent. When the physical age variables are omitted, the estimates decrease to about 1.2 percent. The estimates for mechanical engineers are only slightly lower.

Tables 1 and 2 indicate that the best fit to the data is obtained for profile C, but the differences are very small. In the physical age dimension, salaries increase up to the age of 45 years and decrease thereafter. There are only minor differences in the estimates between the three variants of model I.

In principle, model II states that the investment profile may differ for each age of entry, k_0. We have estimated the models for three intervals of k_0 corresponding to those who completed schooling at ages younger than the normal age (quick), at the normal age (normal), and at ages older than the normal age (slow).[12]

Coefficients for model II are presented in tables 3 and 4. Table 3 presents separate estimates of the parameters of profiles A and C for the three groups of electrical engineers, and table 4 presents the estimates for mechanical engineers.

Stratification of the sample into three groups reduces the degrees of freedom and the variation in the variables, making the estimates less certain. In all cases, however, the covariance tests reported in tables 3 and 4 indicate that stratification is appropriate. The adjusted R^2 is .96 for the entire sample of electrical engineers and .97 for mechanical engineers.

---

[12] The average matriculation age for Swedish engineers varied between 25.9 and 26.5 years during the period 1961–70, with a variance of 3.6–7.8 years. Neither figure exhibits a trend. Accordingly, the data were divided into three groups: the "quick" group, who completed their degrees at or below 23 years of age, the "normal" group, who completed their degrees at ages 24–26 years of age, and the "slow" group, who completed their degrees at or above 27 years of age.

The estimates for profile C generally indicate that salary increases are higher at given levels of experience for the quick group than for the

TABLE 1

ESTIMATES OF ALTERNATIVE INVESTMENT PROFILES FOR SWEDISH ELECTRICAL ENGINEERS FROM MODEL I (STANDARD ERRORS IN PARENTHESES)

PROFILE A LINEAR

Parameter	Equation (14 A)		Excluding Age	
β	.0146	(.0006)	.0122	(.0009)
$\alpha + H_1$	−26.15	(1.25)	−21.31	(1.75)
Experience (j)				
$H_2 + \beta$	.0635	(.0015)	.0896	(.0017)
$H_3 \times 10^3$	−.9848	(.0754)	−1.728	(.0788)
Age (k)				
P_m for k =				
−29 years	.0268	(.0017)		
30–34 years	.0240	(.0017)		
35–39 years	.0164	(.0034)		
40–44 years	.0024	(.0059)		
45–49 years	−.0040	(.0090)		
50–54 years	−.0365	(.0150)		
55– years	−.0082	(.0380)		
$\bar{R}^2$	.9457		.8929	

PROFILE B EXPONENTIAL

Parameter	Equation (14 B)		Excluding Age	
β	.0145	(.0006)	.0121	(.0009)
$\alpha + N_1$	−26.01	(1.25)	−21.29	(1.75)
Experience (j)				
V	.1116		.0079	
N_3	.4255	(.0315)	59.310	(2.707)
$N_2 + \beta$	.0249	(.0019)	−.3603	(.0201)
Age (k)				
P_m for k =				
−29 years	.0249	(.0017)		
30–34 years	.0248	(.0017)		
35–39 years	.0188	(.0034)		
40–44 years	.0021	(.0058)		
45–49 years	−.0076	(.0087)		
50–54 years	−.0369	(.0149)		
55– years	−.0057	(.0138)		
$\bar{R}^2$	.9469		.8932	

TABLE 1 (*Continued*)

PROFILE C: STEP FUNCTION

Parameter	Equation (14.C)		Excluding Age	
β	.0146	(.0006)	.0123	(.0009)
α	−26.08	(1.24)	−21.63	(1.73)
Experience (j):				
$X_m + \beta$ for j =				
1 year	.0491	(.0058)	.0504	(.0082)
2 years	.0689	(.0062)	.0877	(.0085)
3 years	.0708	(.0074)	.0940	(.0104)
4 years	.0759	(.0079)	.0995	(.0110)
5–9 years	.0436	(.0024)	.0646	(.0032)
10–14 years	.0393	(.0043)	.0511	(.0056)
15–19 years	.0314	(.0072)	.0279	(.0096)
20–24 years	.0335	(.0113)	.0252	(.0149)
25– years	.0281	(.0228)	.0080	(.0311)
Age (k)				
P_m for k =				
–29 years	.0234	(.0018)		
30–34 years	.0250	(.0018)		
35–39 years	.0202	(.0035)		
40–44 years	.0101	(.0061)		
45–49 years	−.0082	(.0094)		
50–54 years	−.0372	(.0153)		
55– years	−.0049	(.0377)		
R^2	.9437		.8952	

normal group and are higher for the normal group than for the slow group. This is not true for each experience interval, however, and for several intervals the standard errors are quite large.

IV. Implications

The empirical results provide insights into several of the theoretical issues posed in the previous sections—issues related to "ability," starting salaries, and investment profiles.

First, it is worth pointing out that for both models and for both samples the estimates of profile C suggest that salary increases peak at about 3 or 4 years of experience. The initial rise in salary increments early in the career implies that earnings curves have an inflection point consistent with Becker's original (1962) theoretical insight (but not heretofore observed).

As noted above (assumption 5), both models permit starting salaries to differ by age at graduation. Model I is constrained so that variations in starting salaries depend only on prior labor market experience—previous human capital investment and depreciation. The results of model I indicate that starting salaries increase monotonically with age at graduation, that is, prior net investment is positive This result is illustrated in table 5, which presents estimated starting salaries for the three groups of engineers

In contrast, the more general formulation (model II) permits starting salaries for each group to vary by efficiency in learning (or "ability," as measured by k_0) as well as by prior labor market experience The parameters estimated from the more general model suggest that the starting salary is smaller for the normal group than for the other two groups

TABLE 2

ESTIMATES OF ALTERNATIVE INVESTMENT PROFILES FOR SWEDISH MECHANICAL ENGINEERS FROM MODEL I (STANDARD ERRORS IN PARENTHESES)

PROFILE A LINEAR

Parameter	Equation (14 A)		Excluding Age	
β	0019	(0007)	0103	(0009)
$\alpha + H_1$	−20 93	(1 31)	−17 76	(1 71)
Experience (j)				
$H_2 + \beta$	0611	(0017)	0885	(0016)
$H_3 \times 10^3$	− 9437	(0720)	−1 6313	(0555)
Age (k)				
P_m for k =				
−29 years	0322	(0021)		
30–34 years	0220	(0020)		
35–39 years	0169	(0034)		
40–44 years	0011	(0054)		
45–49 years	0129	(0073)		
50–54 years	− 0317	(0091)		
55– years	0577	(0103)		
R^2	9538		9200	

PROFILE B EXPONENTIAL

Parameter	Equation (14 B)		Excluding Age	
β	0119	(0007)	0104	(0009)
$\alpha + N_1$	−21 03	(1 31)	−17 90	(1 71)
Experience (j)				
V	0806		0304	
N_3	6866	(0504)	5 0416	(1703)
$N_2 + \beta$	0144	(0023)	− 0590	(0039)
Age (k)				
P_m for k =				
−29 years	0293	(0022)		
30–34 years	0225	(0020)		
35–39 years	0195	(0034)		
40–44 years	0022	(0054)		
45–49 years	0125	(0072)		
50–54 years	− 0422	(0088)		
55– years	0581	(0103)		
R^2	9548		9208	

TABLE 2 (*Continued*)

PROFILE C: STEP FUNCTION

Parameter	Equation (14 C)		Excluding Age	
β	0118	(0007)	0101	(0009)
α	−20 63	(1 31)	−17 20	(1 71)
Experience (j)				
1 year	0457	(0077)	0549	(0099)
2 years	0641	(0078)	0871	(0001)
3 years	0778	(0089)	1042	(0115)
4 years	0581	(0089)	0855	(0116)
5–9 years	0450	(0026)	0648	(0030)
10–14 years	0431	(0042)	0554	(0050)
15–19 years	0122	(0068)	0172	(0083)
20–24 years	0430	(0100)	0389	(0123)
25– years	0207	(0150)	− 0005	(0187)
Age (k)				
P_m for k =				
−29 years	0290	(0024)		
30–34 years	0224	(0022)		
35–39 years	0174	(0037)		
40–44 years	0083	(0058)		
45–49 years	0057	(0078)		
50–54 years	− 0425	(0099)		
55–	0582	(0104)		
R^2	9549		9216	

Our analysis suggests that the U-shaped relation between starting salary and age of graduation is the result of two factors—the relatively greater ability of those who are youngest at graduation and the relatively greater labor market experience of those who are oldest at graduation Some indirect evidence supports this interpretation For example, a negative correlation between course grades and elapsed time to matriculation is reported for one Swedish engineering school (Rubenowitz 1962), another study indicates that a major reason for delayed graduation among Swedish engineers is some parallel participation in the labor market (Håstad, Högberg, and Johansson 1974)

Differential access to capital markets could provide an alternative interpretation of these results, especially if those with access to funds also had "connections" leading to better-paying jobs However, public grants and loan guarantees, available to all students, together with tuition-free higher education in Sweden make this interpretation somewhat less plausible It is estimated that 50 percent of Swedish university students subsist exclusively on public grants and loans and that another 36 percent depend "mostly" on such assistance (Statens Offentliga Utredningar 1970, p 20). Another study estimated that only 5 percent of those students who interrupt their graduate careers do so because of financial difficulties (Håstad et al 1974)

TABLE 3

ESTIMATES OF MODEL II FOR ELECTRICAL ENGINEERS (STANDARD ERRORS IN PARENTHESES)

PARAMETERS ESTIMATED	PROFILE A LINEAR INVESTMENT AND DEPRECIATION RATES			PROFILE C STEP FUNCTION		
	Quick	Normal	Slow	Quick	Normal	Slow
$\beta(k_0)$	0171 (0015)	0134 (0007)	0142 (0016)	0166 (0015)	0135 (0007)	0144 (0016)
Intercept	−31 06 (3 00)	−23 77 (1 37)	−25 23 (3 18)	−29 95 (2 90)	−23 89 (1 31)	−25 62 (3 18)
Profile A						
Years	0924 (0031)	0956 (0013)	0895 (0032)			
Years^2	− 0015 (0001)	− 0018 (0001)	− 0020 (0002)			
Profile C						
1 year					0657 (0059)	0553 (0148)
2 years					0927 (0062)	0911 (0166)
3 years				1128 (0142)	1007 (0077)	0778 (0198)
4 years				0947 (0148)	1014 (0084)	0994 (0203)
5–9 years				0744 (0042)	0675 (0026)	0606 (.0053)
10–14 years				0412 (0074)	0566 (0045)	0551 (0102)
15–19 years				0644 (0127)	0318 (0072)	0016 (0194)
20–24 years				0163 (0190)	0264 (0107)	0143 (0343)
25– years				0509 (0296)	− 0216 (0027)	0781 (1434)
$\bar{R}^2$	9437	9833	7660	9480	9848	7661
No of observations	195	319	616	195	319	616
F-ratio for stratification		$F(1,126, 1,118) = 2\ 718$			$F(1,119, 1,099) = 2\ 668$	
$\bar{R}^2$ overall		9606			9607	

TABLE 4

Estimates of Model II for Mechanical Engineers (Standard Error in Parentheses)

Parameters Estimated	Profile A Linear Investment and Depreciation Rates			Profile C Step Function		
	Quick	Normal	Slow	Quick	Normal	Slow
$\beta(k_0)$	0175 (0021)	0126 (0009)	0077 (0015)	0173 (0021)	0124 (0009)	0071 (.0015)
Intercept	−31 89 (4 06)	−22 23 (1 84)	−12 60 (3 01)	−31 36 (4 06)	−21 90 (1 79)	−11.26 (3 03)
Profile A						
Years	0895 (0037)	0945 (0017)	0847 (0028)			
Years^2	− 0012 (0001)	− 0017 (0001)	− 0018 (0001)			
Profile C						
1 year					0707 (0101)	0413 (.0171)
2 years					0956 (0103)	0853 (0180)
3 years				1297 (0256)	1157 (0121)	0768 (.0203)
4 years				0899 (0247)	0864 (.0123)	.0866 (.0199)
5–9 years				0736 (0056)	.0695 (0035)	.0566 (.0047)
10–14 years				0501 (0096)	0560 (.0058)	.0640 (0082)
15–19 years				0571 (0164)	0252 (.0088)	−.0095 (.0153)
20–24 years				0479 (0239)	0393 (0125)	0302 (0252)
25– years				0037 (0312)	0042 (0187)	− 0322 (0423)
$\bar{R}^2$	9276	9739	8416	9295	9756	.8431
No. of observations	179	322	541	179	322	541
F-ratio for stratification		$F(1{,}038, 1{,}040) = 2\ 925$			$F(1{,}030, 1{,}008) = 2\ 918$	
$\bar{R}^2$ overall		9726			9731	

TABLE 5

ESTIMATED STARTING SALARIES BY YEAR OF GRADUATION
(1960 SWEDISH KRONER)

MODEL/GROUP	ELECTRICAL ENGINEERS		MECHANICAL ENGINEERS	
	1961	1970	1961	1970
Model I *				
Quick	1,819	2,125	1,779	2,028
Normal	1,906	2,228	1,886	2,147
Slow	2,045	2,390	2,057	2,344
Model II †				
Quick	2,119	2,525	2,057	2,474
Normal	1,937	2,240	1,898	2,173
Slow	2,140	2,499	2,185	2,378

* Coefficients from tables 1 and 2 for age at graduation $k_0 = 23$ years (quick) $k_0 = 25$ years (normal) and $k_0 = 28$ years (slow)
† Coefficients from tables 3 and 4 for age at graduation $k_0 \leq 23$ years (quick) $24 \leq k_0 \leq 26$ (normal), and $k_0 \geq 27$ years (slow)

Model I also requires that the *relative* differences in starting salaries be the same in each calendar year, while model II does not As tables 3 and 4 show, the estimate of the rate of increase in starting salary is higher for the quick group than for the other two groups This is also highlighted by the starting-salary estimates in table 5 This may reflect a secular increase in the variance of both the distribution of ability and the distribution of human capital accumulated before graduation This phenomenon is a likely result of the expansion of the education system during the period, an expansion which seems to have motivated students with relatively lower efficiency in learning to undertake higher education [13]

The empirical results support the importance of considering both age and years of experience in analyzing returns to human capital investments However, even with these highly disaggregated data, the results provide little basis for distinguishing between a linear representation of net investment ratios and a more complex proportional representation

As we have noted, the recognition that human capital both deteriorates (with elapsed time) and becomes obsolete (with successive vintages) makes it impossible to derive gross investment profiles without indepen-

---

[13] During the period 1955-70, the number of new students admitted to engineering schools expanded from 600 to 3,000, and it appears that the quality of entering students declined For instance, in 1961, 61 percent of those admitted to the program in mechanical engineering at the largest engineering school had the top grade in high school mathematics and 1 percent had the lowest qualifying grade, in 1972, only 7 percent of those admitted had the top grade and 63 percent had the lowest grade (Lindberg et al 1974) Based on these trends, we might expect the proportion of slow graduates to increase There is, however, no such tendency in our data, probably because the share of students with parallel labor market activities declined

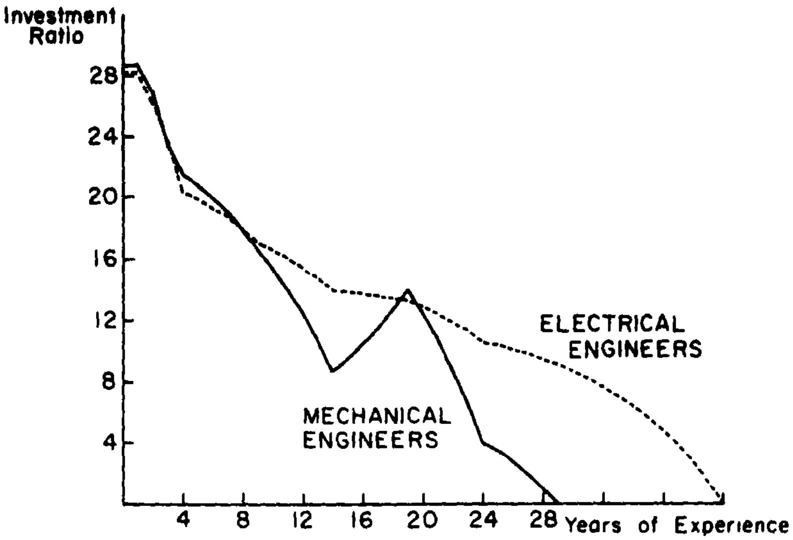


FIG 4 —Estimates of gross investment profiles from model I for workers graduating at age 25 ($r = 13$, $d = 02$)

dent information With a few additional assumptions, however, it is possible to estimate the relationship between gross investment and years of experience Consider, for example, the parameters estimated by profile *C* If we assume that the depreciation rate is constant, knowledge of the retirement age and the rate of return are sufficient to estimate a gross investment profile consistent with any depreciation rate Moreover, since gross investment can never be negative, for any rate of return it is possible, at least in principle, to estimate the minimum rate of obsolescence consistent with nonnegative gross investment

Figure 4 presents such gross investment profiles estimated from model I We assume that for all workers the age of retirement is 65, an independent estimate of the return to education among Swedish engineers is 13 percent (Klevmarken 1972, p 128) The investment profile is drawn for a depreciation rate of 2 percent [14] For any cohort, the age of retirement determines the year of experience at which the gross investment ratio goes to zero, and this allows gross investment in all previous periods to be solved by successive substitution[15] As the figure suggests, if these assumptions are reasonable, the gross investment ratios early in the career are quite substantial Gross investment rates decline rather rapidly throughout the working life, they decline by 50 percent after about 12 years experience

---

[14] Because it was not possible to define steps in the estimated-earnings functions after 34 years of experience, an investigation of the minimum rate of obsolescence is not possible Indeed, for mechanical engineers, the low estimates (essentially zero) for gross investment after 30 years of experience are attributable to a single coefficient (— 0066, table 2) with a large standard error (0256)

Figure 5 shows the gross investment ratios estimated from model II under the same assumptions The figure indicates that the experience profile of gross investment differs among cohorts At any year of experience, the quick group invests more than the normal group and the normal group invests more than the slow group The rate of decrease of investment is also less for the quick group than for the other two cohorts

As indicated in figures 4 and 5, the estimated gross investment ratios increase in the interval containing the first year of experience This finding is consistent with the so-called corner solution implied by Ben-Porath's original analysis (1967), in which it is assumed that the production function for human capital includes the current capital stock as an argument[16]

Although these estimates of gross investment profiles are subject to a generous margin of error, the differences among cohorts are just what we would expect based on the theoretical analysis (see figs 1 and 2)

---

[15] The quantity $c(j)$ is greater than or equal to zero for all years of experience before retirement If the "scrap value" of human capital is realized at retirement (either because a functioning market exists or, more realistically, because the human capital remaining at retirement provides satisfaction), gross investment will continue up to retirement In figs 4 and 5 above, we have assumed that gross investment goes to zero at retirement Thus, knowledge of the retirement age, K, allows us to calculate the years of experience, $J = K - k_0$, for which $c(J) = 0$ for any cohort The other values are found from the expression

$$c(J - 1) = \frac{X_J + c(J) - rd}{(1 + r)} \qquad (6N)$$

If, on the other hand, the "scrap value" of human capital is not realized at retirement, gross investment will cease before retirement In this case, it can be shown that the year of experience, J^*, at which gross investment goes to zero is

$$J^* = J - \frac{\log[1 - (1 - Q)/rQ]}{\log Q}, \qquad Q \neq 1,$$
$$J^* = J - \frac{1}{r}, \qquad Q = 1, \qquad (7N)$$

where $J = K - k_0$, $Q = (1 - d)/(1 + \rho)$, and ρ is the rate of return on riskless securities The qualitative results, in particular the comparisons among the models and cohorts, are similar to those presented above if human capital has no "scrap value" In fact, the reported comparisons among cohorts hold for either assumption for all the rates of return and depreciation rates we investigated ($08 \leq r \leq 15$, $0 \leq d \leq 05$, $05 \leq \rho \leq 15$)

[16] If the initial stock of human capital is small, it is possible that even large amounts of other inputs (e g , all available time) produce small quantities of human capital Under these conditions, output of human capital will increase until an optimum is reached, with input levels less than total available resources After this point, investment ratios decline (see Ben-Porath 1967, Mincer 1974)

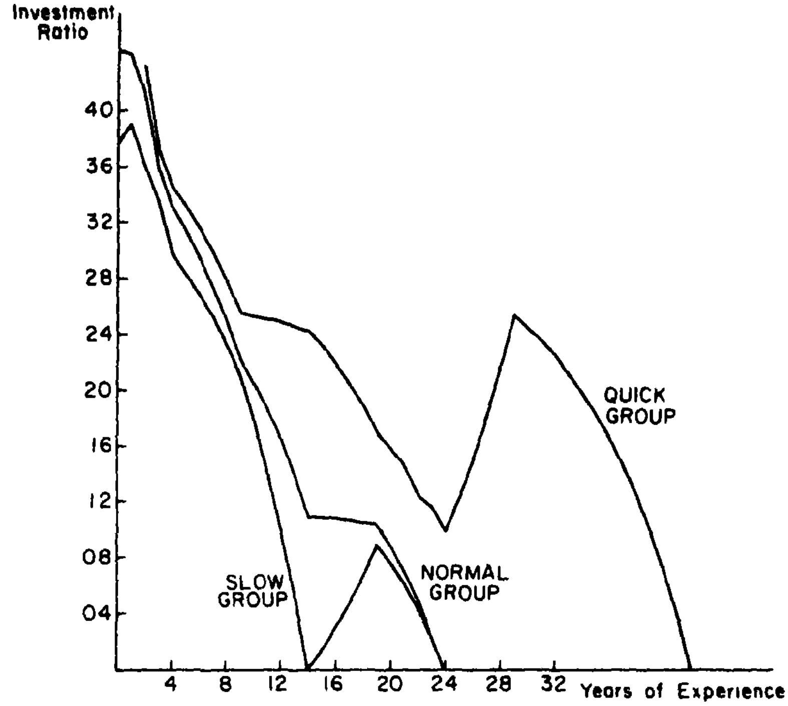


FIG 5—Estimates of gross investment profiles for mechanical engineers for three cohorts (r = 13, d = 02)

younger cohorts with the same education and experience invest more resources in the production of human capital throughout their careers Not only do they invest more absolutely, but their human capital investments as a proportion of potential earnings are higher In terms of calendar age, by age 40 the investment ratio for the quick group is about twice as high as those for the other two groups

These results suggest that age and experience are both important determinants of human capital investment profiles They also suggest that the elapsed time to the completion of graduate training is related to "ability" in ways postulated by the theory of human capital

References

Becker, G S "Investment in Human Capital A Theoretical Analysis " *J.P E.* 70, no 5, pt 2 (October 1962) 9–49

Becker, G S., and Chiswick, B R "Education and the Distribution of Earnings " *A E.R* 56, no. 2 (May 1966) 358–69

Ben-Porath, Y "The Production of Human Capital and the Life Cycle of Earnings " *J.P E* 75, no. 4, pt 1 (August 1967) 265–352

——— "The Production of Human Capital over Time " In *Education, Income, and Human Capital,* edited by W Lee Hansen New York. Nat Bur. Econ. Res , 1970

Davis, J A *Stipends and Spouses* Chicago Univ Chicago Press, 1962.

Flanagan, P , et al "One Dimensional Search Procedures " *Technometrics* 11 (May 1969) 265–84

Harris, S E *A Statistical Portrait of Higher Education* New York McGraw-Hill, 1972

Håstad, M , Hogberg, A , and Johansson, L "Studieavbrott vid de Tekniska Fakulteterna " Mimeographed UKÄ Report no 32 Stockholm Univ Kanslers Ambetet, 1974

Klevmarken, A *Statistical Methods for the Analysis of Earnings Data with Special Application to Salaries in Swedish Industry* Stockholm Indus Inst Econ and Soc Res , 1972

Lindberg, C , et al "Studieeffektivitet vid Teknisk Fakultet " Mimeographed UKÄ Report no 24a Stockholm Univ Kanslers Ambetet, 1974

Mincer, J "The Distribution of Labor Incomes A Survey with Special Reference to the Human Capital Approach " *J Econ Literature* 8, no 1 (March 1970) 1–26

——— *Schooling, Experience, and Earnings* New York Nat Bur Econ Res , 1974

Mincer, J , and Polachek, S "Family Investments in Human Capital: Earnings of Women " *JPE* 82, no 2, pt 2 (March/April 1974) S76–S108

Rubenowitz, S "En Statistisk Analys och Sammanstallning av Data Avsedda att Belysa Studiebegåvning vid Avdelning E, Chalmers Tekniska Hogskola " Mimeographed Gothenberg Psychology Inst , Univ Gothenberg, 1962

Statens Offentliga Utredningar "Behorighet, Meritvardering, Studieprognos Betankande V Avgivet av Kompetensutredningen " Mimeographed Swedish Government Report Stockholm Statens Offentliga Utredningar, 1970

Money Illusion and Balance-of-Payments Adjustment

J C. Murphy and S. K Das

Southern Methodist University

Money illusion, taking the form of lagged price perceptions and viewed in a competitive general equilibrium context, is shown in this paper to reduce the change in a country's balance of payments attributable to a devaluation of given size. Illusion's effect on adjustment is also shown to be probably greater under flexible than under fixed exchange-rate arrangements The results contrast with a view that illusion, in disguising real-income changes, facilitates the elimination of a balance-of-payments deficit

The idea that some form of money illusion is a helpful, perhaps even necessary, element in the correction of an international balance-of-payments disequilibrium has attracted a good deal of attention in the literature on international finance Robert Mundell, in his 1961 classic paper on the theory of optimum currency areas, suggested that the failure of trade unions fully to take into account rising import and export prices is the basis for success in correcting a balance-of-trade deficit by devaluation, and the idea was echoed in McKinnon's 1963 essay on the same topic In two recent papers, W M Corden (1972, 1973) has stated the view in even more unequivocal terms "It is implicit in this analysis that unions do not seek to maintain real wages by forcing a rise in money wages to compensate for the rise in the domestic prices of traded goods Money illusion, or contracts in terms of domestic money, for institutional reasons or convenience, are assumed Hence, money illusion, or labor contracts in domestic money are a prerequisite for the efficacy of exchange rate adjustment" (Corden 1972, p 12) William Fellner (1973), in a paper entitled "Controlled Floating and the Confused Issue of Money Illusion," has recently called the emphasis on illusion in the adjustment process "The New Skepticism "

[*Journal of Political Economy*, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

While there is a good deal of evidence that "illusion" in one form or another exists (see, for example, Branson and Klevorick 1969, Dickson and Starleaf 1972, Dubansky and Kalman 1974), it is disturbing to be told that a sector as important in any nation's market system as the balance of payments is dependent for its self-corrective tendencies on misperception or on behavior that is nonrational in terms of real variables, and one is also led to wonder whether specified illusions work identically in different adjustment processes This paper reexamines some roles that illusion may play in balance-of-payments adjustments (*a*) when correction of a disequilibrium is effected through a change in the exchange rate, and (*b*) when adjustment takes place under fixed rates with price flexibility It begins with a description of an illusion-free correction process that illustrates how a successful adjustment may take place under either regime when prices are flexible Then, with the introduction of money illusion, it shows that this phenomenon affects the adjustment process differently under fixed and under flexible exchange rates It shows also that the effects of illusion, when producers and owners of factors of production are unable to influence their selling price, are generally opposite to those the Mundell-Corden approach takes for granted, it is shown that a lag in perception of general price changes generally impedes the corrective effects of exchange-rate changes It emerges that even where trade unions are present and money illusion serves to mitigate their demands, illusion may not add to the effectiveness with which exchange-rate changes contribute to balance-of-payments adjustment

I. Balance-of-Payments Adjustment without Illusion

It is helpful to view the relevant adjustments within the framework of a general equilibrium model In our image of the process, we envision each actor in the economy as constrained by a budget limitation consisting of an initial endowment of assets plus an income flow for the period A component of each individual's endowment is nominal money, the real worth of which varies inversely with the price level Actors seek, at least in the long run, to maintain money stocks in proportion to their money income and expenditure flows, although the proportions individuals seek to hold may vary with real incomes and with the returns obtainable on other assets Change in the aggregate money supply results from the lending activities of banks and from international reserve flows We assume prices to be flexible and markets competitive [1] For simplicity, we think of the home country as small in relation to the international markets in which it trades

---

[1] Our implicit model is similar to what Harry Johnson (1972, pp 1555–72) has called the "monetary approach" to the analysis of balance-of-payments adjustment except for its abstraction from growth

Consider a single, permanent disturbance to an initial equilibrium in such a system. In particular, assume that the set of goods involved can be dichotomized into those which are tradable internationally (exportable or importable goods) and those which are not (domestic goods), and that at a point in time there is a once-and-for-all shift in tastes in the country away from the set of domestic goods toward the international goods The disturbance results in a tendency toward deficit in the international balance of payments—marked by a rise in the domestic price of foreign currency if the exchange rate is flexible and by a net inflow of goods and a reduction of net domestic financial assets if the exchange rate is fixed

Where the exchange rate is flexible, the shift in demand causes a rise in the home-money prices of international goods as well as a fall in the home-money prices of domestic goods. The rise in the prices of internationally traded goods is, of course, made possible by a depreciation of the exchange value of the currency In the Mundell-Corden approach, wages are resistant to downward pressures, and hence a disturbance of the sort we hypothesize leads to inflation in the home country—partly because prices of international goods rise while prices of domestic goods do not decline, and partly because fiscal and monetary authorities inflate the money supply to avoid unemployment in the domestic-goods sector In our model, characterized by flexible prices, inflation is not inevitable

The disturbing shift in demand is accommodated, when prices and the exchange rate are flexible, by the adjustment of relative prices, much as any demand shift is accommodated in an economy which has no international trade The rise in prices of international goods relative to goods of the domestic sector directs production toward the former and consumption toward the latter, at the same time redistributing real incomes and real-asset stocks among individuals as prices of factors of production and intermediate goods are affected While the initial disturbance is, as Alexander (1952, pp 263–78) and others have noted, an attempt by domestic residents to "absorb" beyond national production, we know that the demand, which would be excessive in the aggregate if exercised, must in fact be deflected by production and consumption substitutions into channels consistent with the overall budget constraint This is a result of the fact that each individual is required to live within the budget constraint comprised of his periodic income and his asset holdings [2] Relative prices in the long run will differ from predisturbance levels less than in the short run as more generous responses to price incentives allow the impact price changes to relax

While the adjustment need not be inflationary, it is worth while noting the possibility that the exchange rate could be a more responsive price than the prices of domestic goods and factors, so that the initial relative price adjustment is achieved more through an increase in international-goods prices than through a decline in the money prices of domestic goods Such an amendment to the model is a step in the direction of the circumstances Mundell and Corden have considered Where the adjustment does result in such an increase in prices on the average, it is clear that an increase in the supply of money and other assets to support spending at this price level must be forthcoming if the higher price level is to be sustained It is clear also that if the monetary expansion is limited, the demand for money to hold will put a limit on the expansion of money spending and hence on the inflation and the depreciation of the currency that transpires

---

[2] Fritz Machlup (1955, pp 255–78) was the first to call Alexander (1952, pp. 263–78) to task for failing to recognize that marginal propensities to consume and invest, and hence "absorb," are functions of relative prices

The discussion of responses to a shift in demand from domestic to international goods applies, of course, symmetrically, with signs reversed, to a disturbance in which demand is transferred from international to domestic goods There, appreciation of the currency on the exchanges allows international-goods prices to fall in domestic-money terms and domestic-goods prices to rise until the disturbing demand shift is suppressed

In the case where exchange rates are *fixed,* the response to an increase in demand for international goods is rather different from the foregoing Since a rise in international-goods prices is here limited by market conditions abroad, the full adjustment required in home-country relative prices must be obtained by a fall in the prices of domestic goods This fall is furthermore accompanied by, and partially enforced by, a transitional loss of domestically earned money income due to an excess of the value of imports over exports and by a discharge to foreigners of domestically held stocks of money balances If foreigners collect reserve money from the banks against these balances, the domestic banks may contribute, by credit contraction, to still further reductions in the domestic money supply In the end, the periodic loss of income and cash balances to abroad ceases as the excess of imports over exports closes, and the closure is enforced by the power of the reduction in prices of domestic goods relative to the largely unchanging prices of international goods to effect production and consumption substitutions The loss of money balances, in equilibrium, holds the general level of money prices down

Again, for a contrary disturbance, when the balance of payments is thrust toward surplus, the responses are reversed

In summary, we can distinguish somewhat different "prime movers" of the adjustment in the flexible and fixed exchange-rate cases When the exchange rate is flexible, taking the disturbance to be a shift in demand toward international goods, behavioral responses by households and businesses are motivated entirely by the increase in money prices of international goods and the fall in money prices of domestic goods There is no necessary increase in aggregate demand for cash balances or meaningful reduction in aggregate real income. When, on the other hand, the exchange rate is fixed, an important stimulus to behavioral responses is the loss by households of money income and money stocks to abroad. In connection with this loss, falling prices (of domestic goods) mitigate the pressures for each household further to curtail real consumption and further to rebuild cash balances, and the price changes, in this way, relieve forces that are leading to balance-of-payments equilibrium. A fall in domestic prices relative to international prices, however, also motivates those production and consumption substitutions which in the end make possible the cessation of international financial-asset flows. Since adjustments to disturbances leading to balance-of-payments surpluses are symmetrical with those leading to deficits, we may say that changing prices energize adjustments under flexible exchange rates but that they de-energize them in some respects while energizing them in others under fixed rates. This generalization is the basis for a number of our comments in the next section on the influences of money illusion.

II. Balance-of-Payments Adjustment with Illusion

Illusion can take as many forms as reality itself. Patinkin's definition of money illusion was very broad: in an initial statement, he made an individual to suffer from money illusion if "his excess demand functions do not depend solely on relative prices and real wealth, inclusive of initial real balances" (Patinkin 1965, p. 22). Money illusion may be rooted in behavior responses as a result of the forms individual utility functions take, but we do not explore that possibility here. Otherwise, it may well stem from information costs (such that people fail to have complete knowledge of price changes and consequent changes in the purchasing-power value of money to respond to) or from expectations (such that people respond not only to prices they perceive in the present but to prices they think to be coming in the future). Since we rule out expectations of future prices being other than present prices, we deal with illusion as if it were due to misperception alone.

When prices in recent periods have not been changing much, it is reasonable to assume that imperfect knowledge about present price changes will lead people to underestimate their extent, households will then overestimate the value of their money incomes and money balances when present prices are generally rising and underestimate the value of this money flow and stock when prices are falling. On the other hand, if there has been recent experience with rapidly changing prices, people with incomplete information could possibly overestimate price changes occurring in this period, in this case, they underestimate the value of money when prices are rising and overestimate it when prices fall.

Furthermore, illusion is likely to be biased. In a society with intensive division of labor, where people are usually sellers of only a few goods but buyers of many, the actors are less likely to be unaware of changes in the money prices of things they sell than they are of changes in prices of things they buy

These comments on illusion lead us to regard it as something which is not categorical but rather as something which is a matter of degree and to classify it as a set of misperceptions which on average either underestimate changes in prices buyers must pay, on the one hand, or overestimate them on the other

We now introduce the effects of some alternative money illusions on the behavioral responses that have been sketched in Section I We rate an illusion as one "impeding" balance-of-payments adjustment if its effect is to reduce the power that real price, income, and money-balance changes have for bringing about reduction in an imbalance in nonmoney imports and exports We rate an illusion as something "contributing" to adjustment, on the other hand, if its effect is to increase the effectiveness of the real market incentives We do not pursue in detail here the influence of illusion on the time paths of variables, although it is clear that dynamic relationships might influence the final result We also avoid specifying final equilibrium conditions and hence the question of whether or not these, too, are modified by illusions that resist being dispelled by the accumulation of experience We are content to show that illusions, in alternative settings, may, at least within important time spans, either emphasize or impair the power of changes in real variables to reduce payments imbalances These effects, for alternative illusions and balance-of-payments disturbances, are summarized in table 1

In case 1, where the exchange rate is flexible, a disturbance that shifts domestic demand onto international goods and leads to a depreciation *of the currency causes money prices of international goods to rise and those* of domestic goods to fall When illusion prevents households and firms from fully perceiving these money-price changes, the result is a reduction in the stimulus the decision makers feel to make the substitutions in production and consumption which will contain the disturbance Larger price changes and a larger depreciation are then necessary, at least in the early phase of adjustment when relative prices make their maximum move, to achieve any given reduction in the potential international imbalance Should depreciation of the currency be accompanied by monetary expansion and general price increases, as in the Mundell-Corden views of things, failure of decision makers fully to perceive these increases and to perceive the concomitant loss in their real incomes and real cash balances would induce delay in absorption curtailment and the rebuilding of real-cash holdings to normal levels Overall spending would then be greater for any set of money prices than it would have been in the absence of illusion, and greater price changes and depreciation of the

TABLE 1

EFFECTS OF ILLUSION ON BALANCE-OF-PAYMENTS ADJUSTMENT

1 Exchange-Rate Regime	2 Balance-of-Payments Disturbance	3 Response Induced	4. Illusion	5 Effect on Balance-of-Payments Adjustment
1 Flexible	Deficit	Currency depreciation	Underestimate of price change	Impedes
2 Flexible	Surplus	Currency appreciation	Underestimate of price change	Impedes
3 Flexible	Deficit	Currency depreciation	Overestimate of price change	Contributes
4 Flexible	Surplus	Currency appreciation	Overestimate of price change	Contributes
5 Fixed	Deficit	Deflation	Underestimate of price change	Impedes or contributes
6 Fixed	Surplus	Inflation	Underestimate of price change	Impedes or contributes
7 Fixed	Deficit	Deflation	Overestimate of price change	Impedes or contributes
8 Fixed	Surplus	Inflation	Overestimate of price change	Impedes or contributes

currency would be necessary to attain temporary balance-of-payments equilibrium For these reasons, table 1 records this circumstance as one in which illusion impedes progress toward adjustment

In case 2, appreciation of the currency allows international-goods prices to fall while domestic-goods prices rise Underestimation of the changes leads again to under response by producers and consumers in making the appropriate substitutions If we think of the exchange rate and international-goods prices as more flexible than domestic prices, the initial change is deflationary Without illusion, a fall in general prices would induce some increase in real spending, financed by increased velocity of money, as spending agents released excess real-money balances With illusion, the gain in real balances and the contribution of lower prices to real incomes goes unperceived, and spending that would have contributed to elimination of a balance-of-payments surplus fails to materialize Here, too, underestimation of price changes impedes adjustment

In cases 3 and 4, where illusion takes the form of overestimates of price change, illusion contributes to the elimination of disequilibrium by accentuating perceived relative price differentials and by accentuating any motivation to restore desired levels of cash balances induced by price-level changes Bias in the perception of the price changes also accentuates average perception of real-income losses on the inflation side and real-income gains on the deflation side Smaller actual changes in the price level and the exchange rate are then necessary to obtain a given degree of progress in the restoration of balance-of-payments equilibrium conditions

In cases 5–8, the exchange rate is fixed In case 5, the disturbance leads to balance-of-payments deficit, and the responses of individuals to the loss of money income and money balances accentuate falling domestic-goods prices We noted at the close of Section I that price changes facilitate this adjustment in some respects but tend to restrain it in others A decline in domestic-goods prices is essential to motivate the consumption and production substitutions which ultimately eliminate the balance-of-payments deficit, and any underestimate of this decline must be expected to slow these adaptations, especially on the consumption side On the other hand, a price decline reduces the motivation of households and firms to rebuild cash balances and to curtail real absorption, thereby relieving a compulsion to adapt In this connection, an underestimate of price declines decelerates spending on all goods, contributing to the reduction of absorption Hence, illusion of this sort may, in net terms, either impede or contribute to an efficient adjustment

For parallel reasons, with signs of the effects adjusted appropriately, the effects of illusions on the other fixed exchange-rate cases are indeterminate Results depend on the relative effects of each illusion on substitution decisions, on the one hand, and decisions in response to tightening income and cash balance constraints on the other

An important point to be noted from the table, then, is that a given illusion does not have symmetric results when a particular balance-of-payments disturbance is corrected (*a*) by an adjustment of the exchange rate or (*b*) with a fixed exchange rate Even in a system of flexible prices, these adjustments are not mirror images of each other, and perceptions of price changes play in each a different role When the exchange rate is flexible, a deficit-caused rise in prices motivates substitution effects and income and cash-balance effects which are reinforcing insofar as reduction of the balance-of-payments deficit is concerned When the elimination of the deficit is obtained through deflation, however, declining prices mitigate the income and cash-balance effects even while inducing the required substitution effects Illusion appears possibly more significant in its net effect on adjustments when the exchange rate is variable than on adjustments where the rate is fixed

Perhaps a more surprising conclusion displayed in the table is the apparent conflict of these results with the Mundell-Corden argument Mundell, Corden, and their followers have argued that failure of trade unions to recognize the implications of rising prices for real incomes and to take steps to offset these setbacks is the secret to a devaluation's success in reducing real incomes and real absorption Table 1, case 1, on the other hand, indicates that failure by spending agents fully to perceive price rises associated with a depreciation of the currency impedes correction of the balance of payments What accounts for these seemingly contradictory conclusions?

The Mundell-Corden position arises from an assumption that households, through monopolistic bargaining, can influence money prices of the labor services they sell Additional assumptions that these trade unions are motivated to maintain the real incomes of all their members in periods of inflation and that the strength of substitution effects is inadequate to prevent them from doing so, lead to the conclusion that money illusion and consequent passivity by trade unions representing sellers of labor in the domestic-goods sector strengthens the power of a devaluation to correct a balance-of-payments deficit.

Our conclusions have been derived on the assumption that all markets are competitive Money illusion of the form that underestimates price changes in our model impairs the effectiveness of a depreciation of the currency The effects we have pointed to are substitution effects (and also cash-balance effects when there is a change in the price level) in addition to the failure of producers of domestic goods to realize their income loss, which is the Mundell-Corden illusion

Associated with this contrast in assumptions about the presence or absence of collective bargaining are contrasts in the implications of the two premises In the Mundell-Corden argument, households which fail to recognize the actual decline in their real incomes consume less than they would have had they been free of illusion and had they demanded compensation In our approach, and assuming some price inflation, all households experiencing illusion consume more than they would have had they simply been illusion free because they accelerate the turnover of money Both propositions are correct

Our conclusions on this score, therefore, seem to limit the practical significance of the Mundell-Corden position on illusion In the presence of trade unions, money illusion may encourage trade-union passivity and in this way lessen inflationary demands But illusion may otherwise exacerbate spending, and it certainly reduces the effectiveness of relative price changes for effectuating the adaptation to new circumstances The net result of a widespread money illusion when trade unions are present is a composite of effects depending on such factors as the relative size of the domestic-goods sector, the strength of the trade unions in that sector, the usual power of cash-balance constraints on spending throughout the economy, the willingness of banks to lend, and the impact of illusion on substitutions

There can be little doubt that balance-of-payments adjustment is possible in a monopoly-free society unencumbered by money illusion And illusion of the sort that fails to recognize the extent of price changes of itself reduces the effectiveness with which given exchange-rate adjustments contribute to restoration of balance-of-payments equilibrium. Combined with trade unions, illusion, on balance, may or may not facilitate an adjustment, but an adjustment in a world subject to illusion is likely to be one in which the early response of the balance of payments to exchange-rate changes is at best disappointing

References

Alexander, S J "Effects of a Devaluation on a Trade Balance " *Internat Monetary Fund Staff Papers* 2 (April 1952) 263–78

Branson, W H , and Klevorick, A K "Money Illusion and the Aggregate Consumption Function " *A E R* 59 (December 1969) 832–49

Corden, W M *Monetary Integration* Publication no 93 Princeton, N J. Internat Finance Sec , Princeton Univ , 1972

——— "The Adjustment Problem " In *European Monetary Unification and Its Meaning for the United States*, edited by L B Krause and W S Salant Washington Brookings Inst , 1973

Dickson, H D , and Starleaf, D R "Polynomial Distributed Lag Structures in the Demand Function for Money " *J Finance* 27 (December 1972) 1035–43

Dubansky, R , and Kalman, P J "The Foundations of Money Illusion in a Neo-classical Micro-monetary Model " *A E R* 64 (March 1974) 115–22

Fellner, W "Controlled Floating and the Confused Issue of Money Illusion " *Banca Nazionale del Lavoro Q Rev* 106 (September 1973) 206–34 Reprinted by American Enterprise Institute as Reprint no 22, February 1974

Johnson, H G "The Monetary Approach to Balance of Payments Theory." *J Financial and Quantitative Analysis* 7 (March 1972) 1555–72

Machlup, Fritz "The Analysis of Devaluation " *A E.R.* 45 (June 1955) 255–78

McKinnon, R I "Optimum Currency Areas " *A E R* 53 (September 1963) 717–24

Mundell, R A "A Theory of Optimum Currency Areas " *A E.R.* 51 (November 1961) 509–17 Paper republished, essentially unchanged, as chap 12 of Mundell's *International Economics* New York Macmillan, 1968

Patinkin, D *Money, Interest and Prices* 2d ed New York Harper & Row, 1965

The Economic Value of Statute Reform: The Case of Liberalized Abortion

---

Timothy A. Deyak
Auburn University

V. Kerry Smith
State University of New York at Binghamton

This paper develops a model for estimating the costs associated with the institutional and legal restrictions related to the states' abortion statutes. A travel-cost demand model is estimated with data for legal abortions obtained in New York State. The purpose is to determine the value—for women seeking abortions—of the 1973 Supreme Court ruling which liberalized abortion codes. The estimated benefits foregone in the absence of liberalization, over a 10-year period with a 10 percent discount rate, range from $12 million to $30 million in present value. It is important to note that this study makes no attempt to measure the losses to certain groups in society as a result of the liberalization.

The legislative process is, in general, subject to considerable inertia. Thus short-run adjustments of existing legal codes to reflect changes in community sentiment must be accomplished almost entirely through variation in the enforcement and/or interpretation of existing statutes by the courts.[1] In many instances these mechanisms are imperfect vehicles for

Thanks are due Kenneth V. Greene, Lex Holmes, Sam Peltzman, and an anonymous referee for most constructive comments on earlier drafts of this paper. The suggestions of Robert Haveman, William G. Johnson, Michael Schooster, and Sharon P. Smith, and the assistance of Timothy Smith of the Office of Biostatistics in obtaining unpublished information on abortions in Upstate clinics by county of origin are also gratefully acknowledged. The usual disclaimer applies.

[1] Stigler (1970) first noted this point. He suggested that "the appropriations to the enforcement agency and the verdicts of juries are the instruments by which the community may constantly review public policy. . . . It is much easier to make continuous marginal adjustments in a policy through the appropriations committee by varying the resources for its enforcement than it is to modify the statute" (pp. 534–36).

[*Journal of Political Economy*, 1976, vol. 84, no. 1]
© 1976 by The University of Chicago. All rights reserved.

altering institutional and legal constraints Since it is generally recognized that these constraints influence economic behavior (see, e g , Buchanan 1962, Roberts 1973), there may be substantial economic losses associated with legislative quiescence The purpose of this paper is to illustrate how these losses might be measured with a recent court decision the reform in the national abortion laws We shall estimate a demand function for legal abortions and measure the benefits as the sum of the consumer surplus increments resulting from the cost savings associated with the reform It should be noted that these benefits accrue only to those individuals who seek abortions

In January 1973, the Supreme Court ruled that a woman may have an abortion performed in any state by a licensed physician as long as it is done within the first 3 months of pregnancy [2] This decision represented for most states (with the exception of New York, and possibly Hawaii and Alaska) an important change in their laws governing induced abortions [3] Over 2 years prior to this decision, New York's induced abortion code was amended to be the most liberal in United States history The New York State Department of Health, which serves as the registrar for information on abortions, has identified the state of origin of those women obtaining abortions in New York prior to the Supreme Court decision Accordingly, these data permit the use of the travel-cost method for estimating demand functions for abortions [4]

Section I outlines our model and relates it to recent developments in the economic modeling of fertility control. The second section discusses the data, our estimates of the demand for abortions, and the measured benefits from the court ruling The last discusses the implications of the analysis for the general problem of the economic costs of institutional constraints.

---

[2] The ruling indicates in addition that states are allowed to regulate but not prohibit abortions during the second 3 months of pregnancy and may only prohibit them during the last 10 weeks

[3] As late as 1972 the laws in 28 states permitted abortion only when the mother's life was in danger In several other cases it was allowed for one or more additional reasons, including rape and incest, or if the child would be born with a physical or mental defect The recent *Newsweek* (March 3, 1975) series of articles on abortion and the law suggests that the effect of the ruling has been spotty, citing the District of Columbia, New York, California, Kansas, and New Mexico as having comparable codes This comparison is somewhat misleading, for it is not the code alone which is important The courts' interpretation is central to the effective liberalization For example, for California Geis (1972) notes that estimates of the impact of the new law was deprecatory Illegal abortions were estimated at 80,000 per year, with tight interpretations of the conditions for legal abortions given as a cause of the continued illegal activity (pp 65–66) Some historical perspective may be useful in understanding the statutes While the opinions on the motivation underlying the legislation differ, it is often argued that at the time these laws were enacted, starting in 1821 with Connecticut, abortions were extremely hazardous Thus, the intent of the law was protection of maternal health Saltman and Zimering (1973) note that a woman's chance of surviving childbirth was 15 times better than the chance of surviving an abortion More recently, Sauer (1974) has argued that, while mother's safety was an important consideration, it would be misleading to count this as the sole objective of the legislation There is equal concern for foetal life The medical profession which was in the forefront of the drive for restrictive laws, stated as their main concern safety of the foetus (pp 57–59)

[4] The travel-cost method has been used extensively to estimate, from individual behavior, the willingness to pay for outdoor recreational experiences (see Cicchetti et al 1972 and Cicchetti, Fisher, and Smith 1973) as well as more topical issues such as the value of the November 1969 March on Washington (Cicchetti et al 1971)

I. Model

The conventional statement of the travel-cost model is in terms of a general gravity model of social interaction and has been widely discussed in the literature [5] However, it can also be considered within a Becker (1965) framework of a consumer producing final services (or basic commodities) In this framework, maximization of a utility function expressed in terms of these basic commodities subject to production, budget, and time constraints will result in a decision rule calling for consideration of the full incremental cost of producing each service flow Prior to the Supreme Court decision and after July 1, 1970 (the date of the change in New York's abortion law), legal abortions could only be secured by a woman transporting herself to New York If we assume that the medical costs and time costs of a legal abortion are fixed regardless of the location in New York where the abortion is performed, then only the travel-cost component of the full service costs changes over a cross section of individual demanders [6] Since these travel costs are related to distance, the distance traveled (adjusted by a price per mile discussed in n 17 below) can serve as a surrogate price for the service

With this rationale, then, we shall estimate a "demand function" for legal abortions and interpret the area under the estimated relationship as a measure of consumer surplus [7] The national liberalization of abortion laws eliminates the need for out-of-state women to travel to New York for legal abortions Accordingly, the components of the full cost which are related to distance are reduced for these women, and there is an increment to their consumer surplus as a result of this price change [8]

---

[5] The essays in Quandt (1970) provide a review of this framework

[6] Time costs can be assumed to have two components One is associated with the surgery and recuperative period and the second the travel time If we assume the first is relatively constant, then the second can be assumed proportional to distance Duffy (1971) cites the range of medical fees for an abortion to be from $145 to $700

[7] Several points should be noted here First we assume that the supply of abortions in New York is perfectly elastic to the individual at a price which is the sum of medical, time, and travel costs Further, we assume that there is no variation in the medical and time costs in the two Upstate New York centers we have designated as the primary sources of the supply Thus costs which can be assumed proportional to distance provide the primary source of the variation in supply price to the individual For a discussion of the importance of these assumptions to the identification of the travel-cost demand function, see Cicchetti et al (1972)

[8] We are not denying that there may also be psychic costs associated with abortion, but we do not have any means of measuring them and have no reason to suspect that they will vary systematically over the points of origin of the women obtaining abortions

The quantity measure used in the travel-cost models is a visit rate That is, the number of visits (or visitors) from each of the origin zones in the sample to the point of interest (in our case the abortion clinics in Upstate New York) relative to a measure of the potential population of users is used as a measure of the expected consumption rate For our model, a logical measure of use would be the number of women coming to Upstate New York for abortions relative to the total number of pregnant women at the point of origin

Before proceeding to discuss the specific application of this model, it is desirable to relate this simple framework to the more detailed analytical treatments of the economics of fertility control Michael (1973) has recently developed such a model to infer the properties of the derived demand for children He casts the decision-making framework in terms of a household production model (extending Becker's 1965 model) and suggests that the fertility-control decision is based on a comparison of the benefit and cost streams attributable to an additional child The net benefit of an additional child is then " the household's effective excess demand for children at the prices, income level, [and the] level of production of complements and substitutes "[9] Michael suggests that education may influence the household's fertility decisions through its effect on any or all of the functions in the model of consumer behavior—namely, the utility function, money-wealth constraint, time constraint, and household production functions If abortion is viewed as one means of fertility control, then it is also reasonable to assume that the education of the individual (or the couple if the decision is made jointly) will be a determinant of the demand for abortions Unfortunately, the indeterminancy in the effects of many socioeconomic factors observed by Michael for the general problem of fertility control is also present in the abortion framework[10] For example, we might expect couples with more education to have greater information on, and therefore easier access to, fertility-control techniques It would then seem reasonable to expect that they will have greater success in preventing unwanted pregnancies Alternatively, education has been recognized to have a strong influence, through its effects on an individual's human capital, on wage rates and earnings It thereby increases the ability to pay for abortions and may make them a more accessible means of fertility control

Additional determinants of the demand for abortion include a measure of income to reflect ability to pay and a variety of demographic characteristics of the population to reflect the effects of tastes and the need for fertility control (e.g., age as a proxy for fecundity) One of the potential demographic factors is religion Michael has suggested (though his data do not provide specific evidence) that Catholics may be more likely to select contraceptive techniques which are less effective and hence have a greater monthly birth probability It may then follow from this specification that Catholics would have greater demand for abortion as their means of fertility control However, this view is in contrast with accumulated evidence on the movement in opposition to legalized abortion Since at least 1869 the majority voice of the Catholic church on abortion has been in accordance with the doctrine of "immediate animation" of the fetus In this statement, Pope Pius IX declared that both early and late nonspontaneous abortion were acts of homicide[11] Sauer's (1974) recent survey of attitudes toward abortion would suggest that the Catholic church has continued to be the most vocal in its opposition to reform in the 1960–73 period He notes· "The plethora of abortion literature in the popular and scientific press that appeared in the 1960's and continued into the 1970's has become increasingly receptive to the idea of complete legalization, except for the Catholic Press "[12] Accordingly, if one accepts both possibilities as potential determinants of the impact of religion on the demand for abortion, then the expected sign of a variable measuring the percentage of Catholics in the origin zone's population is ambiguous

---

[9] Michael 1973, p S132

[10] Ibid , pp S135-S140 The demand functions we shall estimate are probably best viewed as reflecting motives of fertility control Abortions for other motives, i e , women's health endangered, rape, incest, etc are permitted in a large fraction of the state codes prior to the Supreme Court ruling One might suspect that in these cases the transaction costs of obtaining an abortion for one of these reasons in the origin state would be less than the travel costs This presumption may be less true as the proximity to Upstate New York increases Unfortunately, we do not have the ability to discriminate between these cases (to the extent they are in our data) and those for reasons of adjustment in the fertility control decision

A possible alternative to legal abortions are those which are provided illegally It has been suggested that the number of illegal abortions obtained in the United States may be as high as 50–300 per 1,000 live births These estimates are, of course, difficult to obtain and are of questionable value For our purposes, the problem is to determine the price of a reasonable illegal alternative to an out-of-state legal abortion Because of the wide variation in quality, there is a problem in defining a "reasonable" alternative In effect, there is not one illegal market but many, each defined by the health quality of the service provided In addition, information on the availability of illegal abortions may be quite limited and not generally available to the average woman Limited surveys of obstetricians taken prior to the amendment of many state abortion laws indicated that approximately 10 percent admitted referring patients to abortionists [13] Such surveys are, however, subject to problems of revelation and may not be indicative of accessibility to these illegal markets

---

[11] Geis 1972, p 94 Further support for this position is found in several of the Pope's encyclicals For example, as Geis cites, Pope Pius XII in a 1951 encyclical noted, "Any direct abortion whatsoever, even if it is performed in the presence of manifest therapeutic indication to save the mother by its means, when otherwise she together with the child would perish, is immoral and forbidden by divine law Innocent human life, in whatsoever condition it is found, is immune, from the very first moment of its existence, to any direct, deliberate attack "

[12] Sauer 1974, p 64 *Newsweek*'s March 3, 1975 article "Abortion and the Law" cites an NBC News poll which found 58 percent of the voters questioned and 46 percent of the Catholics approved of laws permitting abortion during the first three months of pregnancy or wanted them liberalized These results, if representative, would seem to ameliorate the evidence cited from the Catholic church's position However, support for the law's existence does not necessarily imply a corresponding proportion of Catholic women will consider abortion a viable means of fertility control

Since it is reasonable to assume that a subset of the alternatives available in illegal markets provides important substitutes for legal abortions, some account must be taken of their influence Accordingly, we have reviewed the state appellate court decisions on abortion cases over a period of 24 years and have developed a price series for illegal abortions In most cases we were able to find only one price for each state, and table 1 presents the prices for 29 states We have reported the price cited in the case record, the price in 1971 dollars, the name of the case, and the date The average price paid was $346 in 1971 dollars

There are a number of reasons to believe that these prices are biased [14] Since health complications (often death) appeared to be the most common reason for detecting the illegal abortion, we suspect that these prices relate to the lower-quality strata of the illegal market Moreover, they may have little relation to the equilibrium price for a single level of that market However, since these data are the only information available on prices in the illegal market, we shall include them in the analysis which follows [15]

In summary, we have hypothesized that the abortion rate will be related to the education, income, religious background, age, price of illegal abortions, and measures of the demographic characteristics of the origin point and its population A variety of measures of the socioeconomic factors have been examined In what follows we shall report those relationships which appear to fit the data best, where our criteria of goodness of fit are the $\bar{R}^2$ and the statistical significance of our estimated coefficients

II. Data and Results

The data on abortions are the number of women receiving abortions in Upstate New York after traveling from each of 47 states and the District of Columbia (Hawaii and Alaska are excluded) during the 1971 calendar year While abortions were also performed in New York City, the reports

---

[13] See Laurence Lader (1966, p 32) for more details

[14] We need to also assume that there have been uniform levels of enforcement of the abortion laws over the period involved in our various price estimates Adjustment for the effects of changes in the general price level is, then, only one of a set of alterations which would be necessary to make these price series more representative Unfortunately, it is not clear how such additional adjustments should proceed Enforcement expenditures are not allocated by type of offense so that it is difficult to detect changes in their distribution

[15] We are grateful to William G Johnson for suggesting this approach

TABLE 1

ILLEGAL PRICES FOR ABORTIONS*

State	Price Stated in Record ($)	Price in 1971, Dollars†	Case	Date
Kansas‡	400	400	*State* vs *Darling*	1971
Missouri	125	125	*Missouri* vs *Baker*	1971
Massachusetts	600	600	*Commonwealth* vs *Brunelle*	1971
Kentucky	100	111	*Brown* vs *Commonwealth*	1969
Washington	350	408	*State* vs *Goddard*	1968
Wisconsin	450	524	*State* vs *Gresens*	1968
Tennessee	25	30	*Webster* vs *State*	1967
Connecticut	390	473	*State* vs *Orsini*	1967
Illinois	100	121	*People* vs *Fulton*	1967
Michigan§	500	607	*People* vs *Wellman*	1967
California	650	812	*People* vs *Root*	1966
Virginia	500	625	*Commonwealth* vs *Russo*	1966
Florida	1000	1249	*Rodriquez* vs *Florida*	1966
New Jersey	400	500	*State* vs *Colmer*	1966
New Mexico	125	161	*State* vs *Gutierrez*	1965
Nevada	500	642	*Adams* vs *State*	1965
Texas	50	66	*Catching* vs *State*	1963
North Carolina	150	201	*State* vs *Mitchner*	1962
Pennsylvania	200	271	*Commonwealth* vs *Campbell*	1961
Mississippi	25	34	*Phillips* vs *Mississippi*	1960
Arizona	300	454	*Arizona* vs *Roozer*	1955
Delaware	150	227	*Scott* vs *State*	1955
Indiana	350	528	*Grecu* vs *State*	1954
Idaho	150	226	*Idaho* vs *Rose*	1954
Iowa	50	76	*State* vs *Synder*	1953
West Virginia	75	117	*State* vs *Evans*	1951
Georgia	65	109	*Biegun* vs *State*	1950
Arkansas	50	85	*McClure* vs *State*	1949
Oregon	150	253	*State* vs *Jeannet*	1948

* These cases are indexed in *West's General Digest Fourth Series* 1969 1970 1971 and 1972 and for preceding years the *Sixth* and *Seventh Decennial Digests* Detailed summaries of the cases are reported in seven regional reviews (i e , Atlantic, South, Southeast, Pacific Northeast Northwest and Southwest)

† The consumer price index was used to convert to 1971 dollars

‡ Darling was apparently involved in an earlier prosecution (State vs Darling 1966) in which the reported price was $120 His conviction apparently caused him to increase his prices

§ One other case was found for Michigan in which a price was reported This case reported abortions free to college students (Commonwealth vs Schaflander 1972) Since there were non price-rationing criteria (i e college student), it was not considered representative of the illegal market generally available

of the city's Department of Health do not identify the point of origin of those women coming from out of state for abortions Thus they cannot be taken into account in estimating the demand relationship However, we shall attempt to account for them in our estimates of the increment to an aggregate consumer surplus measure As we noted earlier, the New York State law was changed on July 1, 1970 We have assumed the abortions performed in the six months prior to January 1, 1971 reflect, to some extent, the diffusion of information on the liberalized code and therefore are not representative of the demand patterns for legal abortions

Distance was taken from the Rand-McNally *Road Atlas*[16] We have specified a range of travel costs per mile based on different modes of travel to the Upstate abortion clinics In all cases the travel is interstate, and we have assumed that the costs are proportional to distance traveled These values range from 2 7¢ per person per mile for train travel to 6 8¢ per person per mile for air travel[17] The data on education, income, and age were obtained from the 1972 *County and City Data Book* and the *Statistical Abstract of the United States*, 1973[18] Information on the proportion of the population Catholic was obtained from the National Council of Churches' 1956 report Since records of the number of pregnant women in each state were not available, we have utilized the number of live births in each state as our proxy for the potential population This measure is admittedly imperfect, but appeared to perform better as a proxy for the relevant population than did the number of women in the childbearing ages in each state The Appendix reports the abortion rates (i e , abortions performed in Upstate New York for out-of-state women relative to live births in the state of origin) for each state in our sample, ranked by travel cost

---

[16] Distance was measured from the largest city in each state to Buffalo or Westchester County, the two primary centers for the abortion activity in Upstate New York We then constructed our distance series by selecting the minimum of these two distances

Table 2 reports a sample of our estimates of linear demand equations and table 3 the log-linear equations The travel-cost measure used in these estimates was 5 5¢ per mile times the one-way distance Adjustments to allow for alternative cost per mile measures and/or to account for round-trip mileage require scaling of the travel-cost coefficient in the linear specification so that the product of the coefficient and the newly scaled cost remain constant They do not require reestimation of the relationships[19] In the case of the log-linear equations, scaling does not affect the estimated elasticities but will affect the intercepts Our analysis will focus on the linear specification since it can be a first-order approximation of an arbitrary demand function and simplifies the calculation of consumer surplus The log-linear estimates are presented for comparative purposes Age, income, and education measures were found to be intercorrelated, and therefore it was difficult to isolate a separate

---

[17] Our estimates are based on the following sources (1) automobile, *Costs of Operating an Automobile* (Department of Transportation 1972) 5 5¢ per vehicle mile, (2) bus travel, *Bus Facts* (National Association of Motor Bus Owners 1971) 3 8¢ per mile, (3) train, *Transportation Facts and Trends* (Transportation Association of America 1972) 2 7¢ per mile, (4) air travel, *Air Carrier Financial Statistics* (CAB 1971) and *Air Carrier Traffic Statistics* (CAB 1971) 6 8¢ per mile

[18] U S Bureau of the Census 1973*a*, 1973*b* A variety of measures of education and income were tested We considered the percentage of the population completing high school, the percentage of females completing high school, and the percentage of the population with at least 4 years of college Our income measures included per capita income, median income, and the percentage of families with income in various income brackets We report results only for the final estimates

[19] See Goldberger 1964, pp 185–91 Given the linear relationship is used in benefit estimation, it need not be reestimated for each travel-cost assumption Moreover, since the interest with the log-linear specification is with the estimated elasticities, we do not present the estimates for each assumed travel cost

TABLE 2

ESTIMATED TRAVEL-COST DEMAND EQUATIONS FOR ABORTIONS: LINEAR FORM[*]

INDEPENDENT VARIABLE[†]	EQUATION 1	2	3	4
Travel cost	−9 073 (−3 149)	−10 396 (−3 668)	−9 708 (−3 589)	−12 768 (−4 235)
Four years' college (%) .	0 713 (0 995)			
Median age	0.867 (1 253)			
Urbanized (%)	−0 270 (−2 549)		−0 184 (−2 202)	
Income under $3,000 (%)	−0 086 (−0 238)	0 025 (0 077)		−0 594 (−2 077)
Catholic (%)	0 423 (4 107)	0 381 (3 577)	0 440 (4 665)	
Illegal price	0 007 (1 226)	0 006 (0 952)	0 010 (1 885)	
Intercept	−9 431	5 007	14 506	21 350
$\bar{R}^2$	503	455	510	319

[*] The values in parentheses are Student t-statistics calculated for the null hypothesis of no association, $\bar{R}^2$ is the coefficient of determination adjusted for degrees of freedom

[†] The other independent variables examined are summarized in n 18

TABLE 3

ESTIMATED TRAVEL-COST DEMAND EQUATIONS FOR ABORTIONS LOG-LINEAR FORM

INDEPENDENT VARIABLE	EQUATION 1	2	3	4
Travel cost	−1 042 (−3 518)	−1 422 (−4 565)	−1 443 (−5 102)	−1 140 (−3 981)
Four years' college (%)	−3 233 (−1 844)			−4 695 (−3 156)
Median age	7 985 (1 892)			
Urbanized (%)	−1 281 (−0 939)		−1 684 (−1 371)	
Income under $3,000 (%)	−1 816 (−1 482)	−0 199 (−0 165)		−1 687 (−1 751)
Catholic (%)	−0 284 (−1 081)	−0 258 (−0 909)	−0 165 (−0 634)	
Illegal price	0 077 (0 184)	0 024 (0 053)	0 234 (0 606)	
Intercept	−9 230	0 479	5 579	14 741
$\bar{R}^2$	453	.310	339	440

NOTE — The values in parentheses are Student t-statistics calculated for the null hypothesis of no association, $\bar{R}^2$ is the coefficient of determination adjusted for degrees of freedom effect for each Equation (3) was taken as our "best" model for the linear specification Both travel cost and illegal price were found to be statistically significant with one-tailed tests[20] An income measure did not appear to influence the relationship, however, the percentage of urbanization variable showed a significant negative relationship It should be noted that when the percentage Catholic, percentage urbanized, and the illegal price are omitted from the equation, one income measure does appear to have a statistically significant effect on the demand This influence cannot be attributed solely to income since it may be reflecting the omitted variables' influences on the abortion rate

The results generally conform to our a priori expectations A surprising result was the estimated effect for the percentage Catholic variable Despite the potential for ambiguity in the sign, we expected the effect of Catholic opposition to abortion to dominate, yielding a negative coefficient However, our measure of the percentage Catholic has a positive and statistically significant effect on the abortion rate There are a variety of explanations which might be posed to rationalize this result One might suggest that Catholics use abortion rather than contraceptives as their mechanism for fertility control While the Catholic view on contraceptive techniques is generally one of opposition (with the exception of the rhythm method), this opposition appears to be less inclusive than that associated with abortion The evidence cited earlier on the Catholic church's attitude toward abortion would seem to confirm this view Accordingly, we feel that this explanation is not likely to be correct

A second potential explanation, consistent with the Catholic church's opposition to abortion, would suggest that the greater the Catholic population in a given state, the more stringent is the enforcement of the antiabortion statutes This enforcement increases the risk of performing illegal abortions and therefore increases the supply prices for all quality levels Thus a positive effect for the percentage Catholic variable reflects the restriction of illegal market opportunities This hypothesis would imply that our illegal price measure indicates the effects of one class of substitutes while the Catholic measure captures a wider range of substitutes

Finally, one might suggest that this result is a statistical artifact and does not reflect the "true" relationship A review of the estimated effect for this variable in the log-linear equations indicates that the percentage Catholic has, in all cases, a negative, statistically insignificant effect This result is the only contradiction in the sign of an estimated effect across the alternative specifications

---

[20] As table 1 indicates, there are only 29 of our 48 observations for which we have measures of the illegal price Accordingly, in estimating the model we face a problem of missing observations We have used the zero-order regression method (see Afifi and Elashoff 1967) which replaces the missing values with the mean value over the 29 observations In the case of the linear model (with a single regressor), this method provides unbiased estimates of the parameter but does not improve efficiency over that estimator which used a sample restricted to complete observations

There are likely to be additional explanations for this finding Unfortunately, without additional information we cannot discriminate between them While our tendency is to favor the last explanation, we shall utilize specification (3) for the estimation of consumer surplus because of the strength of the percentage Catholic variable Equation (3) also provided more conservative estimates of the consumer surplus increment than did the others we examined

Before proceeding to discuss the estimates of the consumer surplus gains from the Supreme Court ruling, some attention should be given to the rather unique character of our data and its potential influence on the results First, one might reasonably suspect that individuals traveling from great distances would go to New York City rather than to Upstate New York because it is more accessible [21] While we have attempted to account for these effects between the Upstate centers, we could not include New York City nonresident abortions in our sample, because (as we noted) the point of origin was not identified Table A1 reports the abortion to live birth ratios for the 48 origins in our sample These data indicate that the abortion rate is low for distant states, but it is not clear that we can make an unambiguous judgment on whether our sample is biased toward greater sensitivity to travel-cost changes with this information alone In addition, we have a second problem which may be influencing our sample estimates The data do not include those women with quite small travel costs because they are, for the most part, residents of New York State This omission may serve to distort the demand in the opposite direction [22] These effects may give rise to an errors in variables problem If we assume that the measurement error and the true price measure are uncorrelated, then the bias in our estimates of the price effect will depend on the variance in this error relative to the variance in the true price As this ratio declines, the bias goes to zero In general this view of the problem would suggest that the estimated price effect would understate the true effect

In order to assess the potential bias and to determine the overall representativeness of our findings, we compared the predictions from equation (3) in table 2 with the 1972 abortion rates in Upstate clinics for women from each of the 57 Upstate counties in New York That is, using the estimated travel costs for each county as well as the relevant socioeconomic characteristics for the county, except the illegal price and percentage Catholic which were not available at the county level, we forecasted the abortion rates implied by the model and compared them

---

[21] We are grateful to an anonymous referee for bringing this point to our attention

[22] Robert Haveman brought this point to our attention

TABLE 4

Forecasting Ability of the Model

Statistic	Estimate
Theil's inequality coefficient (U)	966
Partial coefficient due to bias (U_m)	171
Partial coefficient due to variance (U_s)	818
Partial coefficient due to covariance (U_c)	011

Note —See Theil (1961) pp 31-38 for discussion of these measures of inequality of predictions and actual outcomes U_j designates proportion of inequality due to jth effect, hence $U_m + U_s + U_c = 1$

with the actual rates [23] To measure the agreement between the predicted and actual rates, we report Theil's (1961) coefficient of inequality and partial coefficients of inequality in table 4 Theil's inequality coefficient ranges from zero to unity, with zero indicating perfect forecasting and one indicating maximum inequality The coefficient suggests that our model is not particularly good for forecasting However, in decomposing the inequality coefficient into components due to unequal central tendency or bias, unequal variation, and imperfect covariation, we find a low value for the partial coefficient due to bias, indicating that only a small component of total inequality observed is due to this type of systematic error. The partial coefficient due to variance suggests that there are great variations in the abortion rates which our model does not pick up The omitted variables may account for this Alternatively, the character of the within-state demand for abortions may be quite different from that of other states The early liberalization of New York's abortion statutes would seem to imply that this is a reasonable possibility In general, the county abortion rates are larger than those which the model predicts, and the regression coefficient of the actual on the predicted values is 13 96 Accordingly, the model appears conservative in its predictions and is thus likely to understate the estimated consumer surplus

The measurement of the gain as a consequence of the Supreme Court decision requires several simplifying assumptions First, we shall assume that this decision offered all women who were potentially interested in obtaining an abortion an effective supply of the service in their state within at least 180 miles of their homes [24] Second, we shall assume the ruling did not affect any other component of costs (i e , medical or time

---

[23] In forecasting the abortion rates, we used the mean values for illegal price and proportion Catholic This selection was arbitrary and based on minimizing the variance in the forecast error (see Goldberger 1964, pp 210–11)

[24] Clearly the specification of the travel cost after liberalization is largely arbitrary One hundred and eighty miles was selected so that we might approximate the new travel cost as $10 each way, assuming a 5 5¢ per mile cost figure To the extent the new cost is less than this and we expect that it will be in several states, we have underestimated the consumer surplus gain from those states

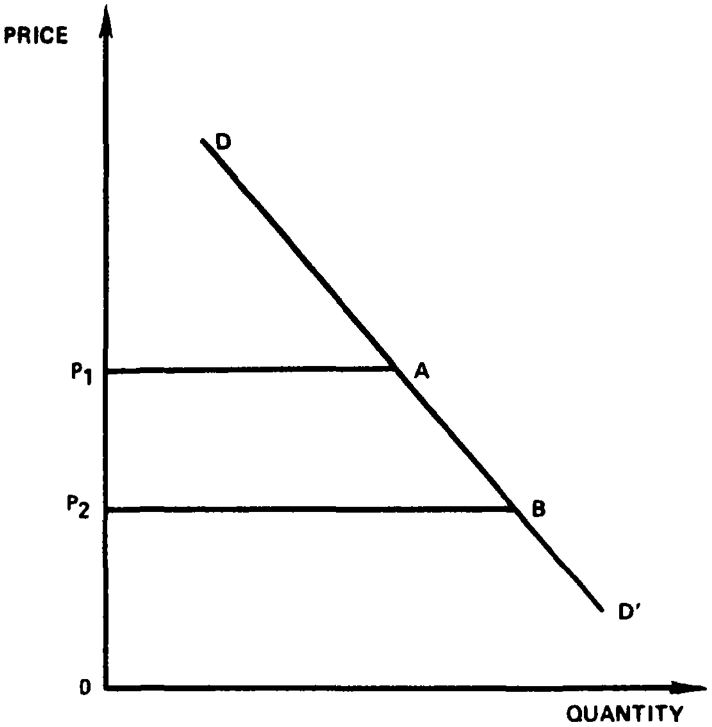


Fig 1

expenses are assumed constant) Finally, the quality of the care in each state after the liberalization will be assumed to be such that all states' services can be considered perfect substitutes Our measure of the increment to aggregate consumer surplus is simply the sum of the increments to consumer surplus in each state In terms of figure 1, DD' is an example of our linear demand schedule with P_1 the price at the old travel cost and P_2 the price at the new The area P_1ABP_2 is the increment to consumer surplus from the price change from OP_1 to OP_2 More formally,

$$\Delta CS = \sum_{i=1}^{n} \int_{P_{2i}}^{P_{1i}} f_i(P)dP,$$

where ΔCS = change in consumer surplus, P = price, and n = number of states Assuming the demand schedules for each state are the same as our estimated relationship, we can calculate the increment to consumer surplus Table 5 reports the predicted total number of abortions after the price change They were estimated by using the demand relationship for Upstate nonresident abortions and the number of nonresident abortions in New York City in 1971

TABLE 5

ANNUAL BENEFITS FROM ABORTION LIBERALIZATION

GEOGRAPHIC BASE OF ESTIMATES	ESTIMATED ABORTIONS	CONSUMER SURPLUS INCREMENT BY TRAVEL MODE ($ MILLION)	
		Train	Air
Upstate New York	14,490*	0.19	0.47
New York City	130,802†	1.69	4.25
Total	145,292	1.88	4.72

* This estimate relates to the number of abortions for other than New York State residents if the price was based on a $10 (each way) travel cost. During 1971 26,698 abortions were performed in Upstate New York for nonresidents. *Our model predicted substantially fewer at the existing travel costs—7,246.* If the actual number of abortions in 1971 is used rather than the predicted and we attribute the average consumer surplus increment to each the total benefits in millions of dollars become 2.03 for train and 5.13 for air.

† This estimate corresponds to the number of induced abortions performed in New York City in 1971 for nonresidents. The prediction assumes the price change has no effect on their abortions. Clearly this is unlikely. However we have no basis for an alternative estimate.

In these computations, we have focused on the implications of the price change for nonresident abortions. Since New York State's abortion code did not require alteration after the Supreme Court ruling, we have assumed New York residents were not affected.[25] Our estimates suggest that the average round-trip savings (consumer surplus) ranges from $13 to $33 per woman. Applying these average consumer surplus increments for Upstate abortions to all abortions in the state indicates that the estimated annual benefits of the price change resulting from the ruling, to those who previously would have traveled to New York for abortions, ranges from $1.9 million to $4.7 million, depending on the assumed travel cost per mile. It should be noted that in all cases we have selected the most conservative estimates of the benefits. Accordingly, there is good reason to expect that the annual benefits will be considerably higher than our estimates. Nonetheless, if we consider the estimated benefits foregone in the absence of the liberalization, for a 10-year period, using a 10 percent discount rate, the present value of these costs ranges from $12 million to $30 million.

III. Concluding Remarks

In discussing an extension to the conventional Pareto allocation criteria, Roberts (1973) has recently concluded that "institutions and similar social parameters are properly a component for the efficient allocation of resources in promoting social welfare."[26] This paper has illustrated a means of measuring the costs of one such institutional constraint—the statutes governing induced abortions Our findings indicate that if existing statutes were to remain in force for 10 years, the welfare losses to those seeking legal abortions might range in value from $12 million to $30 million in present value terms

---

[25] There is, of course, the loss of earnings and employment effects associated with a reduction in the number of abortions. However, the health sector is probably best characterized as one facing excess demand, and accordingly, the losses here are likely to be minimal.

Ordinarily, one might conclude such an analysis with mention that this study is not a full benefit-cost analysis of liberalization in abortion codes, there are losses to certain groups in the community as a result of the decision. The emergence of groups opposed to abortion reform, such as the Right-to-Life organization, indicates that there is an important subset of the population which experiences significant psychic losses from any liberalization in the codes In principle, they ought to be willing to bribe others not to reform Thus we have no evidence, in Kaldor-Hicks terms, on the magnitude of the net benefits (or losses) associated with the change However, given the strong sentiment on both sides of this issue, we feel there is a need to go further and be particularly dogmatic on the use and interpretation of these estimates Our consumer surplus measure assumes that a liberalized abortion code exists in a particular location. The analysis then attempts to measure the benefits to those seeking legal abortions from extending the code generally We do *not* presume to have measured the benefits from legal abortion [27] We have proposed a method for evaluating some of the welfare losses associated with restrictive legal statutes and therefore indirectly suggest that the issues raised by Buchanan and Roberts may be quite significant

[26] Roberts 1973, p 399 He further suggests that "social structures cannot be ignored by being taken as either given or without historical basis Finally, the extent and costs of institutional change often provide incentives for individuals not to reveal their preferences, as in the case of purely public goods " As we noted in n 3 above, the historical view of these codes suggests that when the medical and surgical technology was limited they did serve to prevent welfare losses However, the inertia in legislative action and nonrevelation problems noted by Roberts are likely to have been important in maintaining the codes so long after serving their original purpose

[27] The current interest in Rawls's (1971) contract theory of justice by economists suggests that aspects of moral philosophy may again play a role in some economic analysis There are significant moral issues associated with liberalization of the abortion codes Since our analysis applies to statute reform and is not a benefit-cost analysis of abortion, we do not feel it is appropriate for us to raise these moral issues in these concluding comments

Appendix

TABLE A1

ABORTIONS IN UPSTATE NEW YORK RELATIVE TO LIVE BIRTHS FOR SAMPLE ORIGINS

Origin	Abortion Rate	Origin	Abortion Rate
Oregon	0.06	Georgia	10.13
California	0.09	Missouri	12.53
Washington	0.76	Tennessee	10.47
Nevada	0.60	South Carolina	2.57
Idaho	0.21	Wisconsin	4.69
Arizona	0.33	Illinois	7.21
Montana	0.50	Kentucky	14.08
Utah	0.11	North Carolina	2.09
New Mexico	0.00	Indiana	10.48
Colorado	6.32	West Virginia	4.72
Wyoming	5.67	Maine	17.76
North Dakota	1.55	Virginia	3.09
Texas	1.46	Michigan	9.40
Oklahoma	1.64	New Hampshire	21.85
South Dakota	2.18	District of Columbia	1.28
Louisiana	1.46	Vermont	47.38
Kansas	0.76	Maryland	3.16
Florida	8.29	Ohio	12.18
Mississippi	2.56	Massachusetts	48.24
Arkansas	2.43	Rhode Island	21.27
Nebraska	6.31	Connecticut	22.53
Alabama	1.95	Delaware	2.40
Minnesota	3.56	Pennsylvania	10.07
Iowa	2.07	New Jersey	15.78

NOTE.—The states are ranked according to distance from the abortion center. The abortion rate is the abortions performed in Upstate New York relative to live births (in thousands).

References

Afifi, A. A., and Elashoff, R. M. "Missing Observations in Multivariate Statistics. II. Point Estimation in Simple Linear Regression." *J. American Statis. Assoc.* 62 (March 1967): 10–29.

Becker, G. S. "A Theory of the Allocation of Time." *Econ. J.* 75 (September 1965): 493–517.

Buchanan, J. M. "The Relevance of Pareto Optimality." *J. Conflict Resolution* 6, no. 4 (December 1962): 341–54.

Cicchetti, C. J., Fisher, A. C., and Smith, V. K. "Economic Models and Planning Outdoor Recreation." *Operations Res.* 21, no. 5 (September/October 1973): 1104–13.

Cicchetti, C. J., Freeman, A. M., III, Haveman, R. H.; and Knetsch, J. L. "On the Economics of Mass Demonstrations: A Case Study of the November 1969 March on Washington." *A.E.R.* 61, no. 4 (September 1971): 719–24.

Cicchetti, C. J.; Smith, V. K., Knetsch, J. L., and Patton, R. A. "Recreation Benefit Estimation and Forecasting: Implications of the Identification Problem." *Water Resources Res.* 8, no. 4 (August 1972): 840–50.

Duffy, E. A. *The Effects of Changes in the State Abortion Laws.* Washington: Dept. Health, Education, and Welfare, Government Printing Office, February 1971.

Geis, G. *Not the Law's Business?* Crime and Delinquency Issues Monograph Series, no. 3. Washington: Nat. Inst Mental Health, Government Printing Office, 1972.

Goldberger, A. G *Econometric Theory*. New York: Wiley, 1964.

Lader, L. "The Scandal of Abortion Laws." *New York Times Magazine* (April 25, 1965)

Michael, R T. "Education and the Derived Demand for Children." *J.P E* 81, no 2, pt. 2 (March/April 1973) · S128–S164.

Quandt, R. E , ed *The Demand for Travel. Theory and Measurement*. Lexington, Mass.: Heath, 1970.

Rawls, J *A Theory of Justice*. Cambridge, Mass. · Harvard Univ. Press, 1971.

Roberts, B "An Extension of Optimality Criteria An Axiomatic Approach to Institutional Choice " *J.P.E* 81, no. 2, pt. 1 (March/April 1973) · 386–400.

Saltman, J , and Zimering, S. *Abortion Today*. Springfield, Ill Thomas, 1973.

Sauer, R "Attitudes toward Abortion in America, 1800–1973." *Population Studies* 28, no 1 (March 1974). 53–67.

Stigler, G. J "The Optimum Enforcement of Laws " *J.P E* 78, no 3 (May/June 1970) 526–36

Theil, H *Economic Forecasts and Policy* Amsterdam North-Holland, 1961

U.S Bureau of the Census *1972 County and City Data Book* F Statistical Abstract Supplement Washington · Government Printing Office, 1973 (*a*)

——— *Statistical Abstract of the United States 1973* 94th ed Washington Government Printing Office, 1973 (*b*)

Destabilizing Speculation: A General Equilibrium Approach

Harry G. Johnson

University of Chicago

Friedman's proposition that destabilizing speculators must lose money is easily proved with partial equilibrium analysis But this analysis proves too much In a general equilibrium analysis, assuming speculation implies institutional arrangements for borrowing goods one period and repaying them the next, one party must gain and the other party and the whole community lose from destabilizing speculation, but whether the initiator gains or loses depends on tastes and the commodity denomination of the contract

Ever since Milton Friedman's classic article on "The Case for Flexible Exchange Rates,"[1] theorists have been concerned to dispute one of Friedman's most contentious propositions, namely, that destabilizing speculators must lose money (Note that this does not imply the converse proposition, that stabilizing speculators necessarily make money—see below) The critics have, however, in the judgment of those who have studied the matter, failed to make their case their counterexamples have implicitly either selected arbitrarily a small group of especially clever destabilizing speculators who make money at the expense of other destabilizing speculators (the whole group losing in total), or imputed to destabilizing speculators profits that are not realized—and could not be realized—through market closure of the speculation (see Telser 1959) Yet there remains uneasiness about Friedman's result

Part of the uneasiness stems, obviously, from the emotional attachment

---

I am grateful to my colleagues Lester Telser and Jacob Frenkel, especially the latter, for very useful comments on the first draft of this note

[1] See Friedman (1953), pp 157–203, abridged in Caves and Johnson (1968), chap 25, pp 413–37 See also Johnson (1973), chap 8, pp 198–222, esp p 214, for an analysis similar to Friedman's Friedman's n 8 can be construed as protecting his argument against criticisms implicit in the argument of this note

[*Journal of Political Economy*, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

of many economists to a popular stereotype, applied particularly to the case of speculation of all sorts, that "the speculator" is typically a man who derives a private profit by creating social woe—as characterized most starkly by the speculator who hoards grain in the early stages of a famine (and is sometimes shot for his foresight). But there is a deeper reason for uneasiness, centering on the question of identification and description of "the destabilizing speculator," and specifically from the fact that the unconscious (and apparently not unreasonable) assumptions of partial equilibrium analysis dictate Friedman's counterstereotypical conclusion

The basic difficulty would seem to be that for every buyer there must be a seller, and vice versa, that, consequently, every time that speculation—defined as purchase now for sale later, or vice versa—occurs, there must be (at least) two speculators present, one buying for resale and the other selling for repurchase, that, therefore, one can discuss "destabilizing speculators" and their profits or losses only by concentrating attention on one side of the market in speculative transactions, and that, finally, there is no fundamental rationale for doing this other than the artificial conventions of economic theory itself, and particularly of partial equilibrium analysis with its built-in bias toward looking at the demand side of the market (or at the money side of the money-goods exchange)

The purpose of this note is to analyze the problem in general equilibrium terms, and particularly to show that one group of speculators in a situation in which speculation is "destabilizing" must make profits and the remainder, and the whole group together, make losses In the course of the analysis, some comments are offered on the bias of implicit assumptions commonly held about the nature of a money economy, assumptions incorporated particularly in the Marshallian partial equilibrium, demand-and-supply analysis

The Friedman analysis can, first, be put very simply in terms of demand-supply analysis, as depicted in figure 1*a* and *b*

Figure 1*a* represents a case of pure destabilizing speculation The demand and the supply curve (DD and SS) are constant period by period (and drawn for diagrammatic simplicity with equal and opposite slopes, as are the curves in fig 1*b*) In the absence of speculation, equilibrium is at P_e, with consumers' surplus the area to the left of DD above P_eQ_e and producers' surplus the area to the left of SS below P_eQ_e The speculator now buys MN in period 1, raising price to P_2, increasing producers' surplus by $P_eQ_eNP_2$, and reducing consumers' surplus by $P_eQ_eMP_2$, giving a net gain to nonspeculators of MQ_eN by comparison with the no-speculation situation In the next period he sells $VT = MN$ at price P_1, increasing consumers' surplus by $P_eQ_eTP_1$ and reducing producers' surplus by $P_eQ_eVP_1$, giving a net gain to nonspeculators of VQ_eT by comparison with the no-speculation situation The "destabilizing spec-

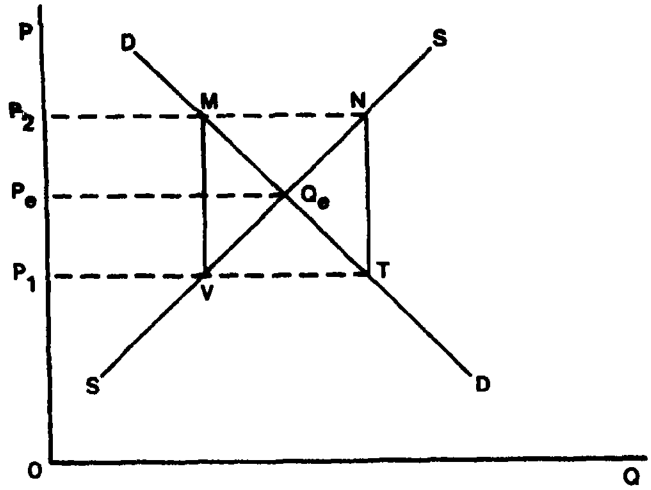


Fig 1*a*

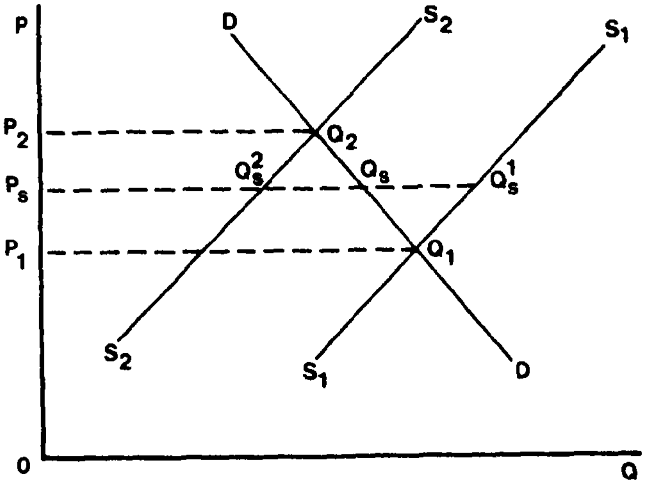


Fig 1*b*

ulator" loses $MVTN$, of which half ($MQ_eN + VQ_eT$) is a transfer to consumers' and producers' surplus and half ($MQ_eV + NQ_eT$) is social loss represented by divergence between the marginal consumption value and marginal cost of output.

Figure 1*b* represents the converse case of perfect stabilizing speculation. The supply curve is assumed to shift over two periods from S_1S_1 to S_2S_2, leading to equilibrium at Q_1 and Q_2 with prices P_1 and P_2 in the absence of speculation. Perfect speculation sets the equilibrium at Q_s with price P_s in both periods, yielding zero profits to the speculator but a net gain of producers' surplus of $Q_1Q_s^1Q_s$ in the first period and a net gain of consumers' surplus of $Q_2Q_s^2Q_s$ in the second period. Note specifically that the perfect stabilizing speculation assumed in the diagrammatic construction leads to neither profit nor loss for the speculators; if speculation were not perfect the speculators could wind up with net profits, if they speculated "too little," and with net losses, if they speculated "too much," while still making the prices more stable (vary less between periods) than would have been the case in the absence of speculation.

Note that we have actually established two conclusions rather than the single one stressed by Friedman: that destabilizing speculators necessarily lose money, and that whether speculators' activities are stabilizing *or* destabilizing, the nonspeculators always gain from the presence of speculation. The second result is even more curious than the first, and, if it is taken seriously, it implies that the whole debate over "destabilizing speculation" is beside the point—as long as we can define "speculators" as people whose welfare we do not care about (and that is implicit in the conventional social connotation of the term "speculator"), the more they speculate, whether in stabilizing or destabilizing fashion, the better for everyone else.

The foregoing remark calls attention to one of the biases in the conventional argument, the identification of speculators with people we do not care about. A more important bias is implicit in, but disguised by, the use of Marshallian partial-equilibrium analysis in the foregoing analysis. It is implicit, in both figures, that the nonspeculative sector of the economy is willing in the first period to supply goods without spending all the proceeds on consumption, and in the second period to consume more than it is currently earning from sales. In short, where the "speculative sector" buys goods now in order to sell them later, presumably in the expectation of a profit, the "nonspeculative sector" sells goods now in order to buy them back later, presumably equally in the expectation of (some sort of) profit. Under these circumstances, how can one reasonably distinguish one group as "speculators" from the remainders as "nonspeculators"? The answer is, only by assuming that the "speculators" are conscious of what they are doing but the "nonspeculators" are not, and building the economic analysis on acceptance of the "nonspeculators" view of themselves. Some apparent plausibility can be given to this procedure by appealing to the nature of "money" as contrasted with "goods" and arguing that the holder of money is by social convention empowered to turn it into goods on demand by purchase, while the holder of goods is not entitled by convention to

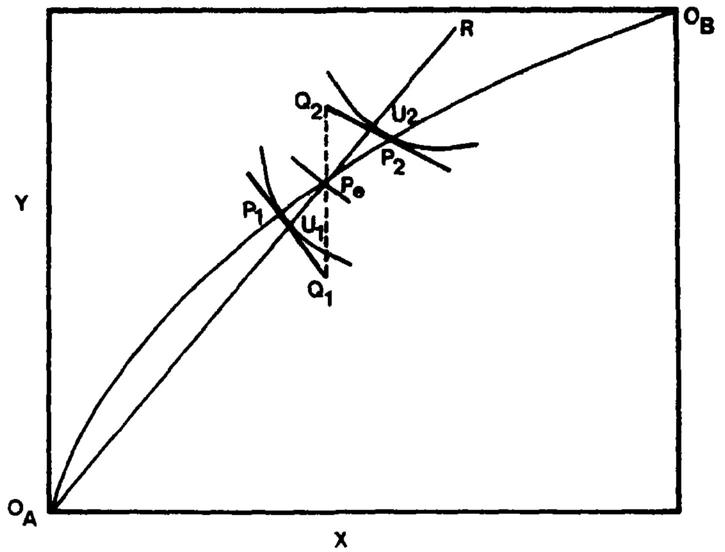


FIG 2

demand that purchasers buy his goods in exchange for money But this argument is inadequate, because the "nonspeculators" are required not merely to sell goods for money on demand but to hold the money and not re-spend it on goods in general and this good in particular All that one can reasonably say on the basis of the analysis presented is that when goods could be produced and consumed in a steady flow over time any activity that causes production and consumption to occur unevenly (and unequally) over time by transferring output between periods reduces welfare, and conversely any activity that transfers goods from periods of abundance to periods of scarcity through storage and de-storage increases welfare, by comparison with a situation in which current supply has to be currently consumed

We now turn to the general equilibrium formulation of the problem, in a "barter" model Figure 2 shows the familiar Edgeworth-Bowley consumption contract box, on the assumption that A has a relative marginal preference for X, the initial (nonspeculative) equilibrium being at P_e Now, suppose that speculation can occur, via an agreement that one can borrow consumption from the other in one period on condition of repayment of the exact amount borrowed in the next—which may be specified in terms of one or other of the commodities or a mixture of them, and for convenience is specified in terms of Y (There is no point in introducing interest, which is an unnecessary complication, since it assumes the presence of profitable investment which will be available whether or not speculation can take place) The two halves of a speculation involve A starting with budget lines through Q_1 and Q_2 $(Q_1P_e = P_eQ_2)$ and trading to points P_1 and P_2 on the contract curve, U_1 and U_2 are the intersections of the relevant indifference curves for A with the ray OR through P_e, distances along which from O represent A's level of welfare on the assumption—reasonable for this problem A and also implicit in the Marshallian technique—that the marginal utility of income to A at constant prices is constant It is evident that the distance U_1P_e is less than the distance P_eU_2, both because the budget line through Q_1 is steeper than that through Q_2 (by a familiar property of the contract curve) and because U_1 and U_2 must lie northwest of the respective budget lines (by concavity of the indifference curves) so that A gains on net from speculation By parallel reasoning it can easily be shown that B must lose by speculation Further, on the assumption that the marginal utility of income is the same for A as for B at P_e—an assumption that is reasonable because otherwise there should be a continuing net transfer of resources between the two individuals sufficient to equate the marginal utilities of income to them and hence maximize social welfare—it can be shown that speculation in these circumstances must reduce total welfare (in other words, be "destabilizing" in the fundamental sense)

We have thus proved that, in the case of destabilizing speculation, one party must lose, the other gain, and the two together lose, the gainer being the one with the relative marginal preference for the *numéraire* good (Note that, on the assumptions made in the previous paragraph, it is only this difference in preference that differentiates the individuals) But who is "the destabilizing speculator"? By the agreement permitting speculation, either party may initiate speculation If we define "the destabilizing speculator" as the party that initiates the speculation, then if that party is individual A he is a "destabilizing speculator" who makes a profit from his destabilizing and social-welfare-reducing actions, but if it is individual B he is a "destabilizing speculator" who makes a loss for himself and for society (but benefits the "nonspeculating" party) Conversely, if the agreement permitting speculation specified exact repayment in terms of X instead of Y, and A initiated destabilizing speculation, he would lose for himself and society as a whole but benefit B, while if B initiated it he would gain for himself but cause losses for A and for society as a whole

It is, in short, a purely arbitrary assumption, though one apparently perfectly natural and reasonable in a money economy, that speculation is defined in terms of equal successive receipts and payments of commodities, not money, that the initiator of speculation is assumed to be the initial holder of money, and that destabilizing speculation involves an initial rise followed by a fall in the relative price of the *numéraire* commodity for the speculative contract—which ensures that Friedman's contention that a destabilizing speculator loses money must hold Note that if it were the holder or supplier of goods who initiated the destabilizing speculation, with the tacit agreement of the initial money holder, the "destabilizing speculator" would gain, though by less than the social loss his actions imposed—as is reflected in the second proposition proved with Marshallian techniques above Further, in the case of speculation in the foreign exchange market, the proximate exchange is of one money for another, not of goods for money, so that one cannot in that case take refuge in the contrasting characteristics of "money" and "goods" to lend credibility to the Friedman result

In view of this analysis, it would seem wisest, particularly given the dominance in the thinking of many economists of conventional stereotypes about the antisocial nature of speculation, to abandon the emphasis of the Friedman analysis on the proposition that "destabilizing speculators must lose money " Instead, attention should be concentrated on the fundamental points that whenever anything can be reallocated over time through storage, the possibility of stabilizing and destabilizing speculation arises, that the use of money, as a virtually costless storable alternative to goods, maximizes the possibilities of both stabilizing and destabilizing speculation, that destabilizing speculation involves a social loss, though generally a welfare gain for some subgroup in the total society, that, consequently, in the particular case of the debate over flexible versus fixed rates of exchange, the empirically important question is whether the speculation inherent in the maintenance of a fixed rate will generate more or less destabilizing speculation than would occur with a floating rate and no (or possibly limited) official speculation on the exchange rate. (A more subtle point, made by Stanley Fischer in another context of discussion, is that the fundamental consideration should be welfare, not price stability, since risk preference may make price variation welfare optimal This point was made earlier by Friedman [1969])

In conclusion, it should be mentioned as a matter of intellectual origins that this note follows directly from work conducted over many years on the economics of reversed international transfers, with special reference to the social consequences of the maintenance of fixed exchange rates (see Johnson [1974], pp 79–110, which includes references to previous literature), and that the most recent piece of this work, which employs a model including two goods and money, with money demand determined by a velocity coefficient, indicates that there is a central range of parameter variation within which both parties to the agreement that makes speculation possible must lose economic welfare as a consequence of "destabilizing speculation" (see Johnson, forthcoming) Finally, the reader's attention should be called to a recent symposium in the *Quarterly Journal of Economics* on the very similar problem of whether the consumer gains from price instability (see Hueth and Schmitz 1972; Oi 1972; Samuelson 1972*a*, 1972*b*, and Waugh 1972)

References

Caves, R F and Johnson H G eds *Readings in International Economics.* Homewood Ill Irwin (for American Econ Assoc, 1968)

Friedman M "The Case for Flexible Exchange Rates" In *Essays in Positive Economics* Chicago Univ Chicago Press, 1953

——— 'In Defense of Destabilizing Speculation" Reprinted as chap 13 in Friedman, *The Optimum Quantity of Money* Chicago Aldine, 1969

Heuth, D, and Schmitz A 'International Trade in Intermediate and Final Goods Some Welfare Implications of Destabilized Prices" *Q J E* 86 (August 1972) 351-65

Johnson, H G "The Case for Flexible Exchange Rates, 1969" In *Further Essays in Monetary Economics* London George Allen & Unwin, 1973

——— "The Welfare Effects of Reversed International Transfers" In *Trade, Growth and Macroeconomics Essays in Honor of Lloyd A Metzler*, edited by George Horwich and Paul A Samuelson New York and London Academic Press, 1974

——— 'The Welfare Economics of Monetary Transfers and Reversed Transfers" Forthcoming

Oi, W Y "Comment" *Q J E* 86 (August 1972) 494-98

Samuelson, P A "The Consumer Does Benefit from Feasible Price Stability" *Q J E* 86 (August 1972) 476-93 (*a*)

——— "Rejoinder" *Q J E* 86 (August 1972) 500-503 (*b*)

Telser, L G "A Theory of Speculation Relating Profitability and Stability" *Rev Econ and Statis* 41 (August 1959) 295-301

Waugh, F W "A Comment" *Q J E* 86 (August 1972) 499

Pareto: A View of the Present through the Past

Vincent J. Tarascio

University of North Carolina, Chapel Hill

This paper examines those factors which caused Pareto to change his philosophical views with age and the effect of these changes on the man and his work. The approach is essentially biographical Finally, the present state of economics is viewed in the context of the historical background of the paper in order to provide a perspective often lacking among economists because of their preoccupation with the present

In his book on Pareto, Professor Finer begins the opening chapter with the succinct observation that "Pareto was born in 1848, the *annus mirabilis* of liberalism He died in 1923, ten months after the fascist March on Rome" (1966, p 3) Other than the fact that Pareto's life span coincided with the decline of an intellectual epoch from the peak of its popularity to its practical demise, such a coincidence would be of little interest Nevertheless, Pareto's own views over his lifetime appear to reflect a change corresponding to the change in the climate of opinion theories of progress, perfectibility, and positivism become the objects of open attack by the end of the nineteenth century The consensus is that Pareto began as a staunch supporter of liberalism and ended as a precursor of fascism, if not its supporter[1] The fact of the matter is that Pareto's *sentiments* remained liberal to his death, although his skepticism caused him to mistrust beliefs, including his own

The purpose of this paper is to examine those factors which caused Pareto to change his philosophical views with age and the effect of these

---

Presidential address delivered at the History of Economics Society Conference, Harvard University, May 23, 1975

[1] Since consensus is often incorrect, one should not be surprised to learn that this particular view is also incorrect In spite of contrary evidence the consensus persists, because such myths assume an existence of their own quite independent of factual realism For more on this issue, see Tarascio (1968, p 11)

[*Journal of Political Economy*, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

changes on the man and his work My approach will be essentially biographical Pareto's life will be divided into three periods—Italy, Lausanne, and Celigny—which represented not only changes in physical location but also changes in orientations Finally, I shall argue that there is a message in Pareto, which is as current as it was in Pareto's time

The Italy Period

Vilfredo Pareto, the son of an Italian nobleman, arrived in Italy in 1858, when he was 10 years of age His father had been living in France in voluntary exile, having been a partisan of the Mazzini movement The extent to which the father's political sentiments influenced the son is not clear It is clear that Pareto possessed a predisposition toward economic liberalism quite early "I was about 16 years of age when I happened to read two authors of contrary nature, namely, Boussuet and Bastiat The former I disliked intensely, the latter fully satisfied by sentiments" (Pareto 1907)

Later Pareto entered the Polytechnic Institute of Turin, where he acquired a professional command of mathematics and where he was graduated with a doctor's degree in engineering in 1868, at the age of 20 By this time, Pareto had become a devoted disciple of liberalism He himself provides a vision of his youthful ideals

> In 1868 I was 20 years of age, and I was able to add reason to those [laissez faire] sentiments It seemed to me then, as it seemed to just about all men, that my convictions were the result of reason I did not realize that my reason was only an attempt to give a logical appearance to that which the sentiment led me to believe
>
> I became enamored of Buckle His work appeared to me to be the ultimate in reason in the social sciences Here I encountered the methods used in physics, methods which I was studying at the University, and it amazed me that there were persons so ignorant and so superstitious as not to understand those doctrines
>
> My creed at that time was approximately as follows Political economy, as it had been constituted by the classical economists, was a perfect science, or almost perfect. All that remained was to put the principles into practice It was necessary to imitate the Cobden League, which was most useful and the most sublime thing in humanity In politics, the sovereignty of the people was axiomatic, liberty was the universal panacea. History showed good, honest, intelligent people oppressed by

> the upper classes, by their own superstition Militarism and religion were the major scourges of mankind. Caesar among the ancients, Napoleon I and Napoleon III among the moderns were malefactors to me. I denied or at least excused the evils of democracy. The Terror was a slight blemish in the luminous picture of the French Revolution [1907]

These were Pareto's beliefs at about the time he embarked on a business career.

After serving as a consulting engineer for the railways, Pareto became general superintendent of the ironworks at Val d'Arno As a businessman, he devoted himself to the practical realization of the liberal ideals; he sided with the friends of liberalism and free trade, he joined and became an active member of the Adam Smith Society in Florence He even stood for Parliament from the district of Pistoria on a free-trade platform He was defeated

In spite of his practical orientations, Pareto's interest in scientific and intellectual fields continued to develop He knew French, English, Latin, and Greek [2] He acquired a knowledge of Greek history and became a learned classical scholar It was during this period that he was introduced to the philosophy of Comte and the economics of Walras The practical, scientific, and intellectual aspects of one's endeavors were not as dichotomized as they are today, and this ambience is reflected in Pareto's decisive friendships with such persons as Domenico Comparetti, the famous philologist and Hellenist, Arturo Linacher, a learned classicist; Sidney Sonnino, the statesman, Giustino Fortunato, the biographer, and A Franchetti, who at the time was working on a translation of Aristophanes They were all members of a group which gathered at the house of Peruzzi, the meeting place of some of the most brilliant minds of the period in Florence The prevailing climate of opinion among many intellectuals was reflected in Pareto's intellectual circle Finer describes this prevailing climate as follows "The age in which he [Pareto] grew up was dominated by belief in social and material progress, in human perfectibility, and in scientific positivism The paths by which thinkers and believers reached these common conclusions were various, but the outcome was much the same In their various ways the fashionable arguments reached a consensus for democracy and rationality, for free trade and the market economy, for pacificism and humanitarianism, and the fact that these arguments were different seemed to make the final consensus all the more convincing" (1966, p 3) What remained was to put these philosophies into practice, a task to which Pareto the reformer devoted himself

---

[2] There is some dispute as to Pareto's knowledge of German For details, see ibid, p 8, n 13

The practical realization of the liberal ideals did not materialize in Italy Indeed, politics in Italy went another way. For Pareto, doubts and disillusion began to appear

> When the left came to power in Italy, when the empire succeeded the republic in France, I was partially forced to recognize that my theories did not correspond to the facts I modified my theories, in part, and drew near to the ideas of a conservative liberal, rather like those of G de Molinari
>
> Studying history closer, I saw that what had earlier seemed simple was instead very complex I began to understand that the word "democracy" signified everything and nothing
>
> In political economy, the nonsense of the opponents always confirmed for me again the economic doctrines of the classicals [Pareto 1907]

With time, he withdrew from the mainstream of Italian politics After the death of his father in 1882, he retired with the intention of preparing himself for a professorship in economics For 11 long years he was unable to secure the position he sought in Italy Both he and Pantaleoni were eventually forced to abandon their homeland and seek employment in Switzerland

Here was a man who, having failed as a liberal reformer, packed up his dreams and carried them to Switzerland in 1893, at the age of 45 Pareto had contributed virtually nothing to science Indeed, even as great an admirer as Bousquet came close to characterizing this period of Pareto's life as a failure

> His immense erudition remained under a bushel He had perhaps not spoiled his life, but nothing suggested that he was to do anything notable in the course of it
>
> He was an idealist, highly intelligent, of indomitable energy, profound honesty, who revolted (rather naively) at seeing the world so little in accordance with his ideal, and who, to the extent that his feeble strength permitted, grappled in vain with the petty realities of political life, and its often sordid coalitions of interests If he had died at forty-two, nobody would have suspected that one day he would be hailed as a thinker, and I doubt whether, even in the most detailed Italian works on economic or political history, his name would even have appeared somewhere in a footnote. This silence would have been fully justified [1961, p 323]

At 45, an age when many persons view their best years as being behind them, Pareto launched himself on a new career as a professor at the University of Lausanne

The Lausanne Period

Pareto arrived at Lausanne in 1893 At first his work in economics reflected the influence of Walras, but with time he tended to drift away from Walras, and his work began to take on its own special flavor As for Pareto's liberal sentiments, these too changed with time Let us begin with his economics. He tells us: "In 1893 when I came to Lausanne I was still a "liberal economist" . The task of teaching compelled me to study it [economics] more thoroughly I saw that many of my theories had to be modified in order to become scientific The same situation occurred later on with regard to sociology when I undertook to teach it Generally, I learned to doubt sentiment, and if something suited my sentiments it became, therefore, only suspect, and I would search for arguments against it with great diligence, as though it were contrary to my sentiments" (1907) The change Pareto describes above is remarkably illustrated in his *Cours* of 1896–97 and the *Manuale* of 1906 In the *Cours,* the theory hardly went beyond Walras's *Elements,* and Pareto lost no opportunity to expound the doctrines of economic liberalism In the *Manuale,* Pareto breaks away from Walras in several important areas of theory, and the economic and social reality is portrayed as it *is* (at least in Pareto's eyes) rather than as it *should be* according to his sentiments Indeed, Pareto uses the occasion to ridicule his own *personal* sentiments, which were still intact at that time, because they could not stand the "test of experience "[3]

In economic theory, his tendency to carry the theory of general economic equilibrium to higher levels of generality reflected his desire to purge economic theory of the ethical connotations of utilitarianism, and also of such concepts as value, capital, prices, etc [4] In their place he substituted "undefined things," subject to physical transformations and transformations in space and time [5] All this to pursue what today is referred to as "neutral observation language" in philosophy

Pareto's "neutral" (scientific) attitude became an obsession, he attacked his former heroes for their failure to maintain an ethically neutral stance Adam Smith and Jean Baptiste Say were excused for their tendency to substitute ethical principles of rationalism for the study of facts, since at that time it had seemed that civilization was undergoing a new birth, and their chief function as economists had been to dissipate ignorance by teaching and preaching the truth (Pareto 1916, 3, sec 2016 1409–10) He was less charitable toward Bastiat, whom he had admired since his youth For the older Pareto (1916, 4, sec 2146, n 7, p 1481),

---

[3] See the important "Proemio" in Pareto (1906, pp v–xii)

[4] For Pareto's own words on the matter of utility and prices, see Pareto (1960, 2 287–93) On capital, see Pareto (1906, pp 283–87)

[5] Apparently Pareto attributed some significance to his conception of the economic process as physical transformation and transformations in space and time (see Pareto 1906, chap 3, 1960, 2 287). For a systematic study of Pareto's conception of transformations, see Tarascio (1973)

Bastiat's work was devoted to justice and right (especially in *Economic Harmonies*), as was Proudhon's work As for Walras, although his works possessed scientific merit they were essentially the expression of a reformer (Pareto 1910, pp 138–39) And so it went.

Pareto's newfound skepticism did not come as a revelation Instead, it grew with his experiences, and one experience which had a profound effect on Pareto's faith in *political* liberalism was the Dreyfus affair in France Let us use Pareto's own words

> At the beginning of the Dreyfus affair in France, I was still a "democrat," and I was still a person who believed it to be possible to govern according to certain principles The conviction of Dreyfus seemed to me the highest point of absurdity and inequity in the circumstances in which it had taken place
>
> The arguments of the anti-Dreyfusards were so silly that they confirmed my concepts even more so Now that I am entirely changed, I have reread the writings of Brunetière, and I find them as false and sophistic as the first time
>
> After the victory of the Dreyfusards, I remained amazed at seeing them use against their adversaries the same wicked cunning they had previously lamented. While some simple minds, such as I, were lost adhering to *principles*, the majority paid attention instead to their *interests* [1907]

Pareto was outraged to see anti-Dreyfusards subjected to injustices similar to those they themselves had previously inflicted upon the Dreyfusards His outrage was not merely cheap, pious moralizing the injustices grated on his sense of fair play This sense of fair play had been demonstrated in Italy, when he opened his house to Italian socialists persecuted by a reactionary government, although he despised their doctrines (Loria 1923, p 431) In the case of the Dreyfusards, it was simply a matter of those not in power pleading on the basis of principles, only to abandon their cherished principles when in power Pareto's dreams of "perfectibility" and "progress" had been all but dissipated With respect to the latter, he observed that "the liberals of one period work to achieve precisely that which they want to avoid" (1907).

More generally, Pareto was reflecting the changes in the social and political environment, events which had not been anticipated in the liberal dream Finer provides a graphic description of the political climate in Europe toward the end of the nineteenth century

> But by the end of the century, theories of progress, perfectibility, and positivism were all open to attack The liberal-democratic compromise was disintegrating into two components which

> were now increasingly regarded as incompatible. . . . In the name of liberalism, the wealthy and the cultivated had denounced state intervention . . ; but the socialists, in the name of democracy, demanded precisely these measures The vector of the two contending forces, the path actually followed by the professedly liberal democratic states, was collectivism. .
>
> The political institutions of the liberal-democratic state were also disintegrating . The French regime was insecure. . .
>
> As for the Italian regime, it was a sordid caricature of all the worst vices of the French. [1966, pp 4–5]

All this was nothing more than an example of the theory in practice One might be disillusioned with the particular application yet continue to subscribe to the theory But the theories of evolution and progress themselves came under attack "But by 1900 the very assumptions on which the philosophies of reason and progress had been erected were under attack The house of Progress had many mansions—in the natural sciences no less than in philosophy From the 1890's onward, voices began to grate out certain things which, if true, demonstrated that the seemingly solid and unshakeable foundations of the mid-century world outlook were so many deceptive quicksands, and the whole gigantic fabric a vast and ponderous mistake" (ibid , p 5) Disappointment is a great stimulus for change, and this was certainly the case for Pareto It was in such an environment that Pareto felt compelled to reexamine the principles upon which his own, as well as the prevailing opinion, had been constructed To do so, he separated himself mentally into two parts Pareto, a normative flesh and blood man, a man of deep passions, seeing to his dismay the liberal dream turned nightmare, and Pareto the scientist, determined to examine social phenomena with an almost ruthless detachment In practice, Pareto vacillated between these positions, confirming the worst of both to those who disagreed with him and the best of both to those who had sentiments similar to his [6]

The Celigny Period

The final period of Pareto's life begins with his retirement from Lausanne in 1906 With the publication of the *Manuale* in 1906, he was able to devote his full attention to sociology Here the problem of ethical neutrality was even greater than in economics He lamented· " 'Humanitarian' sociologies we have had to satiety—they are about the only ones that are being published nowadays Of metaphysical sociologies (with which are to be classed all positive and humanitarian sociologies) we suffer no dearth Christian, Catholic, and similar sociologies we have to some extent" (1916, 1, sec 50 26)

---

[6] This vacillation between Pareto the political polemicist and Pareto the scietnist occurs in the *Trattato*, and it detracts from that work's scientific merits.

Beaten from the houses of economic and political liberals, but not from liberalism, by "reason" and "experience," Pareto embraced those very same objects which had inflicted the pains of doubt and skepticism upon him He set out to "expound a sociology in the fashion of chemistry, physics, and other such sciences" (ibid) His former idols, Comte and Spencer, became examples of practitioners of pseudoscience (Pareto 1902-3, 2 197, 1916, 2, sec 1537 985-86) On the other hand, Marx's concept of "class struggle" was essential for arriving at knowledge of concrete reality, in Pareto's eyes (1916, 2, sec 2020 1412) His appreciation of Marx's concept of "class struggle" was a far cry from *Economic Harmonies*

As if it were not enough for Pareto to cut himself off from his former comrades emotionally, if not theoretically (he borrowed a great deal from them), Pareto isolated himself almost completely from contemporary society by attacking *both* socialist and bourgeois sentiments His earlier two-volume work, *Systèmes socialistes* (1902-3), was such a critical and devastating examination of the history of socialist thought that Hughs (1958, p 78) tells us Lenin feared it more than any other critique In the *Trattato*, Pareto spared no opportunity to attack that caricature of the liberal dream, contemporary European bourgeois society, and to blaspheme the ideological idols of the liberal religion--reason, humanitarianism, democracy, pacificism, etc Ironically, Pareto's attacks and blasphemies occurred at just about the time Europe was plunging headlong into World War I in the name of justice and freedom On the one hand, Pareto's detachment was almost ruthless—his comparison between ancient mythology, which we all perceive clearly, and modern mythology to which we are blind, is very disturbing On the other hand, his obvious ridicule of mankind's stupidity is infuriating, to those not familiar with Pareto's own personal experiences such expressions seem bizarre

All the above seems inconsistent with Pareto's constant reminders that his economics and sociology were constituted on principles of "logico-experimental" science (reminiscent of Marx, who too thought that only he was doing social science) Raymond Aron (1937, p 518) dismissed Pareto's sociology as an immense "deviation" (ideology) What is important for my purposes is that Pareto's new vision of society, the expression of which was difficult for him (as it was later for Keynes in economics), was reflected in his work

Pareto's Paradigm of Society

As early as 1907 Pareto the scientist had extricated himself from his earlier beliefs His new vision of society involved three fundamental principles, which were opposed to those of rationalism and liberalism

One "norm" which he believed very useful for understanding history and sociology is that, "in general, humans act nonlogically, and they believe and in turn want others to believe that such actions are logical" (Pareto 1907) This principle later became the basis for his theory of residues and derivations in his sociology

A second principle involved a distinction between scientific validity and social utility To use Pareto's words "It is important to understand that absurd doctrines can be socially useful, that the various social classes are rather able to have different doctrines, and therefore, what amounts to the same thing, understand the same doctrine differently" (ibid) This distinction between scientific validity and social utility of doctrines represented an important departure from Pareto's former rationalistic stance He says, "At another time, this proposition would have caused the many hairs I had on my head to bristle, and now it leaves the few I have left quiet" (ibid)

As if the principles of nonlogical behavior and the distinction between scientific validity and social utility were not enough of a departure from rational-liberal doctrines, one more principle was added for good measure The last principle called attention to the distinction between the individual and the social, and its implications for policy To use Pareto's words,

> Today we know little or nothing regarding the effects of change on the social state He who advocates such changes does so blindly He knows where society is obliged to go, but he does not know if it will truly go there Therefore perhaps the single norm, in the absence of one better, is that everyone minds his own interests and is content to seek out immediate effects which eare easily foreseeable I say in the absence of better, because not even in this manner is the danger avoided of ending up in the long run where he may not have wanted to go [Ibid]

Pareto, then, stresses the *uncertain* nature of the social process from the point of view of individuals Belief creates certainty, but that was not the kind of world the utilitarians had had in mind It is precisely the existence of this uncertainty which makes individuals so susceptible to ideology What does the slogan "will of the people" mean in such an environment? Certainly not what the liberal-democratic reformers had envisaged [7]

Pareto's sociology, which was begun in 1907, was built upon these "norms" or principles, and the particular theories of residues, derivations, heterogeneity, and class circulation derive from them According to Finer, "It is best to regard Pareto's *Treatise* not as the last of the 'great systems,' but as an anti-system a system to kill all systems " He continues, "Whatever else it did, it killed stone-dead the positivistic-rationalistic types of theories which preceded it and foreclosed all further attempts on those lines" (1966, p 82) Finer is correct when he says that no one attempted "great systems" building after Pareto But to attribute to Pareto the role of "great systems" killer is much too strong Nevertheless, Pareto's sociology did raise some important issues with respect to the social sciences

---

[7] One of the slogans of liberalism was "Educate the masses and everything else will take care of itself," according to Joseph Schumpeter (1954, p 133) Although "education" may reduce some forms of ignorance, it does not eliminate uncertainty Indeed, it may increase uncertainty

To begin with, if human behavior is nonlogical, is not scientific behavior, being a part of human behavior, also nonlogical?[8] Although nonlogical behavior encompassed the greater part of human behavior, Pareto believed there were important exceptions, such as some forms of scientific, economic, political, and military activity (Pareto 1916, 4, sec 2079 1442–44, sec 2255, 1578–82, secs 2400–2411 1728–39, sec 3530, n 1, p 1828) Therefore, Pareto's type of antirationalism did not deprive him of a science of society Indeed, the assumption of rational behavior is unnecessary to science and also superfluous for Pareto His principle of nonlogical behavior does not imply irrational behavior In other words, *both* logical and nonlogical behavior involve consistency of behavior As long as behavior is consistent, generalizations in the form of "laws," such as those of science, are possible Pareto's principle of nonlogical behavior was perfectly compatible with the methods of positive empirical science From current perspectives it is difficult to understand all the fuss about his principle of nonlogical behavior from a scientific viewpoint Certainly the reactions to his sociology by some scientists in the 1930s were extrascientific, to say the least[9]

The second of Pareto's principles, the distinction between validity and social utility of doctrines, would be too much for even Marx to swallow For Marx, like most academic rationalists, believed in the social utility of social science[10] One implication of this principle is that in practical life, truth (validity) is less important, especially in social matters, than scientists and scholars like to believe All this grates on the minds of those who have dedicated their lives to the search for truth, only to be told that their efforts do not really matter much[11]

Keynes, too, seemed to understand the point Pareto was making Recall Keynes's assertion that "practical men, who believe themselves to be quite exempt from any intellectual influences, are usually the slaves of some defunct economist [and] the ideas of economists and political philosophers, both when they are right and when they are wrong, are more powerful than is commonly understood" (1936, p. 838). The reason for the popularity of this oft-quoted assertion among economists is obvious; nevertheless, its more subtle implications have been ignored Keynes was suggesting that ideas are valued for their ideological import irrespective of their scientific validity, which leads one to wonder whether or not economists matter as *scientists*

---

[8] Marx was confronted with this problem when he argued that the "ordinary mind" lives in a world of "illusionism" Nevertheless, Marx distinguished between actors and "true" scientists who are given to the quest for scientific knowledge The classical economists did not fall into the latter category (see Karl Marx 1906, 1 591–92, 1909, 3 198, 199, 369, 1016–17, 1022–23)

[9] For example, William McDougall (1935, pp 36–61) dismissed the *Trattato* as a "crazy book", Max Millikan (1935, pp 324–37) concluded it should be kept from tender young minds

[10] For a detailed study of the methodologies of Marx and Pareto, see Tarascio (1972)

[11] Indeed, one has the impression that for Pareto the *material* progress of mankind has been due largely to the development of the physical sciences and technology, rather than to the efforts of scholars and reformers

Finally, Pareto's third principle, uncertainty of outcome of social change or policies, creates a formidable obstacle between what *is* and what *will be* Granted that science can determine what is (or more accurately what has been), it need not follow that it can predict what will be[12] Marx, the "great systems" builder, was never troubled by this difficulty, since Marx the scientist and Marx the prophet were one

Pareto argued that the interdependence of social phenomena is such that the progress of any one discipline depends upon the progress of others According to Pareto, many of what appear to be exclusively economic phenomena depend on phenomena which are outside the scope of economics (1916, 4, sec 2219 1551) Income distribution, economic growth and development, economic fluctuations, etc , fall into this category Therefore what appear to be random from a purely economic perspective, may not be in the context of a broader scope of analysis For Pareto, sociology was the synthesis of economic, political, psychological, and other factors which together determined the *social* equilibrium

Pareto went one way, toward synthesis, to deal with the problem of interdependence, the social sciences have gone another way, toward greater specialization within and among disciplines Such specialization has allowed for a greater *depth* of understanding, but this greater depth of understanding has occurred at a cost to *breadth* of understanding

Pareto, then, was the last of the "great systems" builders, if one means by the term a synthesis of various aspects of social behavior—economic, political, religious, and so on He was the first of the "great systems" killers, if by "great systems" one means a grand synthesis of positive science and ethics Pareto preserved a place for science in his new vision of society, but the partnership between positive empirical science and liberalism (and other isms) was dissolved Although he acquiesced in this duality (Pareto the scientist and Pareto the liberal), the fun of building a better world on scientific foundations was gone—instead, better-world building involved a struggle of beliefs.

A View of the Present

Today as in Pareto's time, social scientists often see the present only through the present sociocultural influence colors our interpretations of the present, consciously or unconsciously, and has a bearing on the state of social knowledge Therefore, it might be worthwhile to view the present state of economics and the social sciences through the past

---

[12] Certainly the performance of so-called forecasting models in economics seems to bear this out

Now as in Pareto's time theories of progress, perfectibility, and rationalism prevail in the intellectual community Since the 1930s the climate of opinion has changed from one of liberalism to one of interventionism in the United States So now progress, perfectibility, rationalism, and interventionism characterize the prevailing climate of opinion [13] This drift toward interventionism is reminiscent of that which occurred on the Continent around the turn of the century Whether or not the new post-1930s intellectual synthesis in the United States will be more successful in experiencing the practical realization of its dreams than the liberal synthesis of the nineteenth century remains to be seen [14]

The economics profession has experienced an enormous increase in prestige during the postwar period This may be partly due to the great prosperity of the period, in which case we have been bestowed with the dubious honor of receiving credit which we have not deserved [15] It is partly due to the fact that interest in "economic policy," which has become almost an obsession among most economists, reinforces the interventionist climate of opinion In either case, economics (and the social sciences in general) has become the handmaiden of the prevailing climate of opinion, as was the case with orthodox (liberal) economics in Pareto's time

The prevailing obsession with policy or "policy implications" of research may impede the progress of economics as a science [16] In this regard, it might be worthwhile to keep in mind that most so-called economic or social problems are *subjective* in nature, as Pareto stressed in his sociology [17] For example, pollution of the environment has existed for years, but it became a "problem" only when it was perceived as such—and so it was with poverty, discrimination, and so on. Unfortunately, such "problems" are often "solved" in the same way they are discovered, subjectively, which is to say they are often forgotten (after appropriate legislation, of course). Who, for example, worries about the War on Poverty today?

---

[13] I am hesitant to add positivism to this list, although perhaps I should since, as Pareto showed positivism has received more attention in theory than in practice

[14] Signs of disillusion with that modern manifestation of progress - economic growth and development - are appearing Yet "sentiment" of interventionism is becoming stronger

[15] Postwar prosperity is probably due less to "policy" than to demographic, "institutional" (stabilizing influence of defense spending, etc), and other such factors Indeed some writers, such as Friedman, argue that "policy" has been on balance disabling In Friedman's case, this view is consistent with his political philosophy

[16] It appears that economists are in danger of providing a "scientific" veneer for preconceived notions based on beliefs Certainly the existence of so-called conservative, liberal, Democratic, Republican, Marxist, and radical economists is incompatible with a science which views truth as validity

[17] Recall that Keynes described changes in economic activity as deriving from changes in subjective factors such as marginal efficiency of capital, liquidity preference, propensity to consume and save, and so on Nevertheless, these were treated as given by Keynes, so that the subjective basis for fluctuations is never explained Pareto too was interested in economic change as well as social change, but, unlike Keynes, he proceeded to investigate the subjective factors in his sociology

The durability of many conditions observable over long periods of time (such as inequality of income distribution) and the failures of past attempts to alter such conditions should indicate that they derive from factors which go much deeper and which are more fundamental than is commonly understood. Investigation of these conditions and their causes is extremely time consuming and difficult, requiring years of uninterrupted and systematic research. In the public's mind, the realization of the complexity of such "problems" with attempts to "solve" them is no detriment to action. Indeed, the clamoring for action which accompanies the discovery of a "problem" demands action, any action. Whether or not such action is a solution is another question. The logic is perfect: subjective problems are solved subjectively. Of course, I do not wish to convey the impression that social "problems" should be dismissed because they are subjective in nature. Instead, I wish to stress that too much preoccupation with the present tends to distract us from the pursuit of knowledge for its own sake and hence deprives science in the long run of its potential.[18] Indeed, most so-called economic problems are perennial in nature, as historians of economic thought know. One wonders to what extent the problem and policy orientation of the economics profession has dissipated a great deal of precious talent and energy, in order to keep up with ever-changing public sentiments.[19]

Finally, because of the uncertainty of outcome, beliefs remain the basis for most policy—ideology and interests, not science, the dominant forces in social behavior. This, then, is a view of the present through the past, a past which has not changed in many ways in spite of Pareto's efforts—a circumstance which makes his vision of society even more convincing.

References

Aron, R. "La Sociologie de Pareto." *Zeitschrift fur Sozialforschung* 6, no. 3 (1937): 489–521.

Bousquet, G. H. "Vilfredo Pareto (1848–1923)." *Banca Nazionale del Lavoro Q. Rev.* 58 (September 1961): 317–60.

Finer, S. E., ed. *Vilfredo Pareto: Sociological Writings.* New York: Praeger, 1966.

Hughs, S. H. *Consciousness and Society.* New York: Knopf, 1958.

---

[18] It might be worthwhile to recall that Pareto's major contributions to economic science occurred when he abandoned his obsession with contemporary problems and when he reexamined his own present in the context of the past. This perspective provided a basis for his work which was of lasting significance.

[19] Public and private fundings of research tend to be biased toward "problems and policy" orientations.

Keynes, J. M. *The General Theory of Employment, Interest and Money*. New York: Harcourt Brace, 1936.

Loria, A. "Obituary, Vilfredo Pareto" *Econ J* 33 (September 1923): 341.

McDougall, W. "Pareto as a Psychologist." *J. Soc Philosophy* 1 (October 1935): 36–51

Marx, K. *Capital. A Critique of Political Economy*. 3 vols. Vol 1, translated by S. Moore and E Aveling, 1906; vol 2, translated by E. Untermann, 1907; vol. 3, translated by E. Untermann, 1909 Chicago Kerr

Millikan, M "Pareto's Sociology." *Econometrica* 4 (December 1935)· 324–37.

Pareto, V. *Cours d'économie politique*. 2 vols. Lausanne: Librairie de l'Université, 1896–97

——— *Les Systèmes socialistes* 2 vols Paris. Giard et Brière, 1902–3.

———. *Manuale d'economia politica* Milan Società Editrice Libraria, 1906.

——— Letter to Antonio Antonucci, November 24, 1907 In *Alcune lettere di Vilfredo Pareto pubblicate e commentate da A Antonucci* Rome· Maglione, 1938 Reprinted in *Vilfredo Pareto Epistolario, 1890–1923*, edited by Giovanni Busino. 2 vols Rome Accademia Nazionale dei Lincei, 1973

——— "Walras" *Econ J* 20 (March 1910). 138–39.

——— *Trattato di sociologia generale* 4 vols Florence. Barbara, 1916. English translation *The Mind and Society*, translated and edited by A Livingston 4 vols New York Harcourt Brace, 1935

——— *Lettere a Maffeo Pantaleoni* Edited by Gabriele de Rosa. 3 vols Rome· Banca Nazionale del Lavoro, 1960

Schumpeter, J *History of Economic Analysis* New York Oxford Univ. Press, 1954

Tarascio, V. J. *Pareto's Methodological Approach to Economics: A Study in the History of Some Scientific Aspects of Economic Thought*. Chapel Hill Univ North Carolina Press, 1968

——— "Marx and Pareto on Science and History A Comparative Analysis" *Keio Econ Studies* 9, no 2 (1972) 23–35.

——— "Vilfredo Pareto On the Occasion of the Translation of His *Manuel*." *Canadian J Econ* 6 (August 1973) 394–408

Economics in Dialectical Dialect

Martin Bronfenbrenner

Duke University

I

Of the full-dress introductions to economics or political economy that I have yet seen, this one is by all odds the most peculiar and idiosyncratic For such sheer idiosyncracy (in its own day) one should return to von Thünen's *The Isolated State* (1826) or Cournot's *Mathematical Principles of the Theory of Wealth* (1838) (I doubt that Mr Marcus's product will have the staying power of the two works cited for comparison)

In each of the three examples listed immediately above, an author is addressing at an allegedly "principles" level a select audience not normally or professionally interested in economic issues Von Thünen was writing for estate managers, agricultural accountants, and agricultural statisticians of reflective and analytical bent. Cournot was writing for professionals in the mathematical sciences, including physics, astronomy, and engineering along with mathematics proper Marcus is writing for dialectical philosophers and social scientists of Marxian persuasion, whose knowledge of Marxian economics has lagged behind their knowledge of general Marxism, and is, moreover, integrated only incompletely with the remainder of Marxian thought

II

As Marcus's frontispiece we have a "genealogical" diagram of "Principal Currents of Modern Epistemology " Then comes a dedication "To my opponents, who made this book necessary." Next we have a 10-page "Author's Foreword," including a bristling reading list for prospective instructors and also exegeses on topics ranging from Marx's knowledge

Lyn Marcus, *Dialectical Economics An Introduction to Marxist Political Economy* (Lexington, Mass D C Heath Co , 1975) Pp xx + 490

[*Journal of Political Economy*, 1976, vol 84, no 1]

© 1976 by The University of Chicago All rights reserved

of mathematics to the problems of translation between and among English, German, French, Spanish, and Greek The 32-page chapter 1, which now appears, is entitled "The Crisis in Marxian Scholarship", it concerns economics only in passing This "guide" indicates the opacity of the author's organization It may also discourage the casual reader and convince whoever dares venture further that he will be dealing with something very different from the predigested compendium suggested by the word "introduction" in the title

Why so forbidding an introduction? One reason may be Marx's own exemplar, the horrendous 55-page "Commodities" chapter which opens *Capital*, volume 1 Like Marx, Marcus is addressing primarily readers adept at German dialectical philosophy as developed from Kant to Marx via Hegel and Feuerbach His principal task is to apply this dialectic to economics, somewhat as Professor R G D Allen, for example, proposes to apply to economics the corpus of university mathematics (differential and integral calculus, determinants and matrices, differential and difference equations, calculus of variations, set theory, etc). An important difference, however, is that Marcus is addressing an audience whose dialectics already transcends its economics, while Allen's audience comprises mainly economists whose mathematical reach exceeds their grasp

In addition, Marcus's book has an important secondary task, which Allen's books do not, the two tasks continually conflict and block each other with the reader as victim Marcus's secondary task is to convince fellow-Marxists that he (Marcus) understands Marxian dialectical materialism correctly, including its economic applications, whereas numerous other self-styled Marxian economists at various times and places commit egregious errors of understanding and interpretation. The objects of Marcus's contempt include the influential *Monthly Review* group in the United States (Baran, Magdoff, Sweezy), Ernest Mandel (author of the two-volume *Marxist Economic Theory* [1962, English translation, 1968], and anyone he can call a "revisionist " They also include Lenin, Trotsky, Stalin, Mao, Castro, "Soviet economists," and the New Left Even Marx's collaborator Engels, Marcus maintains, fell into occasional error after Marx's death Marx himself is sacrosanct in both "young" and "mature" incarnations Aside from the Master, Marcus's principal hero is a heroine—Rosa Luxemburg, Stalin's major "left-wing deviationist" victim among professional economists, E A Preobrazhensky, also receives a share of his admiration I am ignorant of Marcus's own sectarian affiliation, if any, within the Marxist movement, but he (Marcus) leaves a distinct impression, redolent of the 1930s, of the one-man-party member with whom the world is out of step

The discourse is conducted in occasionally disorganized fashion Its tone is standard Old Left, deliberately contentious and combative It revives the sort of controversy one associates with Marx himself in his "rule or ruin" approach to the First International; such controversy dates back much further, at least to theological diatribes in which one's opponents are "imps of Satan" and their establishments "the Great Whore of Babylon" if not "Sodom and Gomorrah." As in much such writing, Marcus is if anything harsher in his judgment of individual rival Marxists than of individual bourgeois heathen

As regards content, Marcus's book is perhaps 50 percent dialectical philosophy, with a strong epistemological stress The remaining 50 percent appears divided fairly equally between history (including economic history), anthropology-cum-sociology, and economics (including a surprisingly large loading of business administration) Within economics, we find neither systematic presentation of non-Marxist positions, significant confrontation of Marx with his "bourgeois" critics from Böhm-Bawerk to Samuelson, nor development of Marxian theory along lines which blunt these criticisms (such as the "transformation problem" of the relations between values and prices) For a 500-page introduction to economics, in sum, the economics is disappointingly thin

III

Marcus is, then, talking to the Marxian faithful He is obviously interested in "fishing for men," but his fish are the ostensible Marxists whom he finds heretic or renegade There is almost nothing in his economic discussions to entice a bourgeois economist into the fold.

There is, however, considerable philosophical (epistemological, methodological) argument which may accomplish some such purpose Marcus sees conventional economics as a withered arm of philosophy, somewhat as Spaniards and Germans reportedly see Portuguese and Yiddish, respectively, as withered arms of their own languages Philosophy in Great Britain took a wrong track in the seventeenth to eighteenth centuries—"reductionism," as Marcus calls it—under the dominance of empiricists like Locke and particularly Hume. Philosophy was rescued from the Humian dragon by a succession of St Georges from Germany, clad in new dialectical armor—Kant, Hegel, Feuerbach, Marx—but not before Adam Smith, a friend and associate of Hume, already founded bourgeois economics on erroneous predialectical philosophical foundations

Being myself incompetent in philosophy, decidedly including its German-dialectical variants, I shall let Marcus explain for himself why he considers Humian "reductionism" so sinful

> [T]he fundamental fallacy of ordinary understanding is the delusion that the universe is reducible to simple substance, or—the more Hume-like view—that the content of human knowledge is limited to simple-substance-like, self-evident sense perceptions

> This discredited outlook—whether it takes the naive mechanistic [form] or the equivalent mechanistic outlook of empiricism—is termed *reductionism* All varieties of reductionism are formally premised on the fallacious assumption of formal logic, that the universe can be represented as discrete points interconnected by formal relations The intrinsic fallacy is identified by comparing such constructs to attempt[s] to complete a line by asymptotically filling in points In short, reductionism demands *the fallacy of formal induction* [P 15][1]

It appears to follow, although I do not believe Marcus makes the point explicitly, that bourgeois theorists' concern with *prices* represents reductionism compared with Marxian concern with *values* Reductionist fallacy is then allegedly involved in adjusting or modifying a value theory like the labor theory of value to fit more neatly into price theory Economists should have operated in precisely the opposite direction

Since economics deals with epiphenomena like prices rather than with basic matters like values, its search for refuge in formal mathematics compounds the methodological fallacy Marcus, himself, reaches this result by a different route

> Like physical science, economics deals with the relationship between cause and effect However, economics, like all social sciences, also deals with an equally determining subjective process how man's interpretation of nature results in human action The important schools of political economy differ mainly in their expressed or implied explanation of the relationships between the two notions of value thus associated with each commodity
>
> Most political economists implicitly accept the fallacious postulate that the price and utility of commodities are in formal correspondence Marx's economics alone specifies that the two "criteria" are so qualitatively distinct that no mathematical solution of their interconnection is possible If one accepts Marx's analysis, *political economy differs from physical science in that its problems do not admit of systematic formal-mathematical interpretation* [P 33, italics mine]

Another passage illustrates the application of such methodological critiques to economics in practice The subject under discussion is the first Nixon's administration's oscillation between monetarist and Keynesian policies in 1969–71:

---

[1] Italics Marcus's For example, this goes well beyond F H. Knight's concept of "reductionism" which involved only the treatment of more complex matters (e g , human beings) *exclusively* in terms of relations on a lower level of complexity (e.g , atoms or molecules)

> What caused both Friedman and Keynesian schemes to fail, with equally drastic consequences, was that built-in stabilizers cannot work any longer. When the ratio of growth of fictitious accumulation (potential illiquidity) to expansion of production is at a critical stage, any monetary policy—constipative depression, manic diarrhea, or the "safe, sane" in-between approaches—must lead to disaster. Different monetary policies are merely different ways of speeding toward the same inevitable result.
>
> What also fails in such periods is the method employed by empiricist economists. The economic constructs they have adduced from periods of capitalist stability represent correlations which persist so long as the fundamental epistemological flaw in the empiricist method is not reflected empirically. At the extremes of the boom-bust cycle, it must seem to the empiricist that he has encountered a domain of discontinuity, at which point all the laws of the universe have been suspended. What is actually demonstrated during such critical phases is the problem of Praxis. [P. 260][2]

IV

Marcus has apparently had the advantage of more private-business experience than the great majority of academic economists. One can guess that, blackballed by his radicalism from academia and civil service, he has turned to business in something like desperation—and been rather good at it. Certainly the range of his experience, as reported here, is broader than the ordinary economist's—including specifically mine. It is accordingly interesting to notice triple interactions between Marcus's business experience, his philosophical background, and his policy conclusions.

Much of Marcus's business experience, as he reports it, has been at the exploitative frontier of "white-collar crime," bordering on fraud both in the inducement and in the *factum*—a circumstance that I should imagine pushes one to one or the other end of the ideological spectrum

---

[2] Marcus had previously (p. 258) referred to "the so-called Chicago School of Milton Friedman and his co-workers" as "the most wretchedly naive school of thought in theoretical economics." Friedman and his supporters "guided by obsessive belief in the wisdom and efficiency of the invisible hand, rationalized their religious belief in such a deity by extracting from statistical studies a certain rate of growth of the money supply over the long pull of capitalist development." They then, "adduced, *post hoc ergo propter hoc*, that governmental 'interference' ought to be limited to maintaining the indicated optimal rate of growth of the money supply." An original reinterpretation of the *Monetary History of the United States*!

Marcus's experience extends to the speculative overcapitalization of capital values, creating "fictitious capitals" which cannot later justify themselves by earning capacity in the normal course of events Observation of the overcapitalization process confirms Marcus in an overcapitalization theory of depression, of the sort associated in America with Thorstein Veblen—but Marx came first Furthermore, Marcus follows him (Marx, not Veblen) to the point of arguing that recovery "depends upon massive credit from a foreign economy that is not a depressed economy either massive credit from a foreign economy that is not depressed or massive looting of wealth in the form of primitive accumulation from outside the productive resources of the depressed capitalist sector itself" (p 400) [3] Marcus has also been involved with industrial engineering and management science, including "rule-of-thumb" decision rules which appear to have soured him against bourgeois econometrics generally and reinforced his methodological biases On the technical side again, Marcus claims to argue from the inside as well as the outside that "the U S economy, viewed with some knowledge of the ABC's of technology, is one horrendous mess of waste, redundancy, obsolescence, and managerial incompetence" (p 221) Likewise, he believes that the living standard of the representative fully employed U S worker has fallen since the end of World War II [4]

What seems most comment worthy about all this is *not* its paucity of supporting evidence and its immunity from disconfirmation—Marcus sets no store bv logical positivist criteria of meaning, and we would be "talking past each other" to argue along the lines of a methodology he does not accept More comment worthy, rather, is the fewness of Marxists and other radicals bred by and in our jungle of small and middle-sized American business at the border of legitimacy—not to mention the crude materialism of the representative American college of business administration My tentative suspicion is that the George F Babbitt businessman is rescued from dangerous thoughts by some combination of having no thoughts at all—banished by deadly routine plus sheer stupidity—by a certain low-grade chronic sadism toward the consuming public and the public welfare, and by exaggerated notions of the horrors involved in socialist austerity

---

[3] Marcus continues, along Rosa Luxemburg's line "The foreign capitalist may wish to buy up idle man-power and other capacities at the prevailing bargain rates—as the U S bought Europe and other sectors at bargain prices at the end of the Second World War Or, mobilization for looting other countries or massive, previously untapped natural resources would provide the general alternative to foreign aid "

[4] This statement is qualified in evasive ways, making it difficult to disconfirm I quote from p 389 "In sum, if the market basket is defined on the basis of the quality of material culture needed to produce family members qualified for the mode of employment represented by the most modern general technology, the value-content of the standard of living of most trade-union members in the United States significantly declined from 1948 on—a secular decline in the level of real income as determined by any scientific measure of the *necessary* content of the market basket "

V

Marcus recaptures neither the confidence of the Communist Manifesto in a socialist future the day after tomorrow nor the confidence of the pre-1914 Social Democrat that socialism in itself would be a step forward whenever and wherever it came His vision of the immediate future in America is of fascism not significantly more appetizing than the Nazi variety His vision of socialism however, includes along with social ownership of the means of production a regime where most consumer essentials are free public goods, and where their production is planned rationally with an input-output model of the microeconomy grafted to an essentially Marxian macroeconomy of "extended reproduction " All this, however, is sketched only vaguely, as with Marx, the main thrust of Marcus's book is the criticism of capitalism

But what of freedom in Marcus's thinking and in his brand of socialism? He appears to be what my late Wisconsin colleague Selig Perlman called an "efficiency intellectual " This is to say, Marcus believes all rational men of goodwill accept his own technocratic design for the planned economy with minimal need for repression All this is relatively standard, and Marcus also accepts a standard radical view that freedom today (in economic matters) is less than we believe it is because we have been narcotized by advertising, by salesmanship, and by the planned unavailability of goods we really want at prices we really can afford In addition, like many other dialectic philosophers, Marcus sets off freedom against necessity in a fashion well adapted to rationalize almost any measure of dictatorship Admitting that freedom requires the recognition of necessity, who is to draw the institutional frontier between the two domains?[5] Judging perhaps unfairly from his controversial manner, Marcus impresses at least one reader as a Me-for-Dictator type to whom it would be dangerous to entrust the task of drawing any boundary between the domain of freedom and that of necessity or order

VI

I have an impossible dream about an ideal "introduction to economics " It should present both orthodoxy and a goodly sample of the various doctrines comprising "the other economics " The author should present rival positions so favorably—poisoning no wells—that the reader should judge him a Marshallian from his discussion of Marshallian economics, a Marxist from his discussion of Marxism, a Keynesian from his discussion of Keynes, and so on

---

[5] I personally accept this particular dialectic proposition only in a temporal and psychological sense One cannot emotionally savor freedom without previous knowledge and experience of oppression—or oppression without previous knowledge and experience of freedom

After faulting standard English-language principles books for ignoring and neglecting "the other economics" I have sought for utopia among the dissenters, only to find wells poisoned left and (more particularly) right The volume under review is unfortunately a case in point Perhaps I have been dreaming my dream in the wrong place Might not some magnanimous and tolerant conventional economist cognizant of "the other economics" be easier to find than a dissenter tolerant and magnanimous enough to understand and spend time on "the standard stuff" as well as his own recipes for salvation?

Communications

Some Evidence on the Factor Content of West Germany's Foreign Trade

Robert M Stern

University of Michigan

A multifaceted explanation of the determinants of international trade in manufactures seems reasonably well established, based on works by Hufbauer (1970), Baldwin (1971), Branson (1971, 1973), Branson and Junz (1971), Leamer (1974), and Morrall (1972)[1] It is nevertheless useful to examine additional empirical evidence relating to the composition of trade and, in the process, seek to clarify some issues that have not been completely resolved to date Thus, in the present paper, we analyze some evidence on the factor services embodied in West Germany's foreign trade in 1962 and 1969 and consider such issues as (1) the scaling of variables, (2) the aggregation of physical and human capital, (3) whether changes in the relationships have occurred between the years indicated, (4) whether there are differences vis-à-vis the industrial countries and LDCs, (5) how the results may be affected by the exclusion from the analysis of natural-resource and certain other industries, and (6) the influence of nominal and effective tariffs

---

Helpful comments on earlier versions of this paper were received from Alan V. Deardorff and members of the Research Seminar in International Economics at the University of Michigan, Moheb Ghali and members of the Faculty Seminar at the University of Hawaii, Gerhard Fels, and an anonymous referee Jan Kmenta, E Philip Howrey, and Saul H. Hymans provided advice on computational procedures Gerhard Fels was kind enough to supply me with the original source materials used in his own work that permitted construction of the data used in the present study Computational assistance was rendered particularly by Christopher F Baum and by Willa Ann Cohen and Jonathan Francis

[1] A review of these and other studies of the determinants of trade is contained in Stern (1975*b*).

[*Journal of Political Economy*, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

I. The Model

Our basic point of departure is a three-input neo–Heckscher-Ohlin model in which an industry's position as a net exporter or net importer is related to the factor services derived from its physical capital, human capital, and unskilled labor Since we will not observe factor services embodied in the trade of other countries, this model cannot be tested in a logically complete manner, and any findings reached will not necessarily validate or invalidate the model as such Yet if the results of the empirical tests as applied to a single country seem plausible, we may have discovered something worth knowing Thus, in this spirit, our model can be written symbolically as

$$(X - M)_i = f(K_i, H_i, L_i) \qquad i = 1 \quad n, \tag{1}$$

where $X - M$ is the trade balance of industry i, and K, H, and L refer to the industry stocks of physical capital, human capital, and unskilled labor

The model can be expanded to take into account other phenomena besides the neo–Heckscher-Ohlin ones just indicated Thus, on the basis of what we have learned from the product-cycle-technological gap models and from other more realistic models, we could have a more comprehensive model

$$(X - M)_i = g(K_i, H_i, L_i, SK_i, \mathrm{R\ \&\ D}_i, t_i, ES_i, FTD_i), \tag{2}$$

in which variations in industry trade balances may reflect such technological phenomena as the inputs of technically skilled workers (SK_i) and expenditures on research and development (R & D_i), tariffs (t_i), economies of scale (ES_i), and innovativeness as approximated by the first trade date (FTD_i) when a new product appeared

We have some a priori notions on how well endowed particular countries are with certain productive factors and how technically progressive or protective they may be But since we commonly observe only one country at a time, it is difficult to establish what the expected signs of all these variables in equation (2) may be[2] The lack of information is also a great handicap for a single country, so that empirical investigation will perforce be circumscribed, as we shall see momentarily

II. Empirical Specification and Results

Our choice of West Germany as the focus for study was dictated by the data available in research by Fels (1972, 1973)[3] His studies were designed

---

[2] See Katrak (1973) for an attempt to explain the relative export performance of the United States and United Kingdom for selected years in the 1960s in terms of the relative industry characteristics based on human skills, R & D, and scale economies

[3] In an earlier study based upon trade and input-output data for 1954, Roskamp and McMeekin (1968) concluded that West German exports were especially intensive in human capital Our results noted below are consistent with theirs, although our human-capital variables were measured differently and our data classifications do not correspond

TABLE 1

ESTIMATES OF REGRESSION EQUATIONS EXPLAINING
WEST GERMAN NET EXPORTS FOR 1962 AND 1969

EQUATION	CONSTANT TERM	INDEPENDENT VARIABLES *K*	*H*	*L*	*N*	R^2	*F*
1962	−585 73	− 148***	0 473***	− 093	25	780	24 749**
	(1 50)	(3 369)	(6 027)	(055)			
1969	−1492 30*	− 153**	3 032***	3 897	25	777	24 398***
	(1 89)	(2 556)	(5.214)	(1 330)			

NOTE —Here and in subsequent tables, the *t*-statistics are shown in parentheses with signs omitted
* Significant at 10% level
** Significant at 5% level
*** Significant at 1% level

to examine the structure of West German industry with the object of identifying those sectors that might give way competitively and the adjustment problems that might be encountered if production in and exports from LDCs were to be expanded significantly Since Fels's major concern was different from ours, we shall not dwell on any detailed comparison of results

Data were available for 25 industries for 1962 and 1969 [4] Exports were measured free on board (fob) in current Deutsche Marks (DM) and imports were measured at cost plus insurance and freight (cif) These trade data were available on a regional basis as well as in total The estimates of physical capital refer to plant and equipment measured in DM prices of 1962 Human capital was approximated by capitalizing at a 10 percent rate the difference between each industry's wage and salary bill and the amount applicable to male unskilled wage earners Unskilled labor was represented by total industry employment A skill index was constructed by taking the 1961 ratio of engineers and technicians to total industry employment The only other variable for which information was available was industry tariffs, both nominal and effective, for 1970

We first ran ordinary-least-squares (OLS) regressions for the three-input Heckscher-Ohlin (H-O) model, taking the absolute levels of each variable The results are shown in table 1 It is evident in both years that net exports were negatively related to physical capital, positively related to human capital, and that the employment variable was not statistically significant in either period These results are basically similar to those obtained for 1964 U S trade by Branson (1973), except that he found the employment variable to be statistically significant and negatively related to net exports The difference in the signs and coefficients of the physical capital and human capital variables indicates that these measures of capital cannot be readily aggregated as several authors have done[5] It is evident, finally, that there was a marked increase in the coefficient on human capital between 1962 and 1969

---

[4] The data sources are listed in Fels (1972) and will not be repeated here The detailed data used are available upon request from me

The foregoing results should be interpreted with care since they may be influenced by variations in industry size That is, industries that are large in absolute size in terms of factor inputs may have correspondingly large absolute trade balances To correct for this, the industry trade balances were divided by the value of gross production for each industry in the 2 years[6] We also divided the physical capital and human capital variables by the total number of employees The resultant capital/labor ratios seem more in accordance with the way in which the H-O model is usually formulated

We then introduced a variable designed to measure the availability of technical skills in terms of the ratio of engineers and technicians to total employment, $(E + T)/L$, in each industry This skill ratio was available only for 1961, but was used for both 1962 and 1969 in the absence of better data The reason for introducing the skill ratio separately is to test whether it reflects technological influences that cannot be distinguished from or that may exist apart from the industry stocks of human capital

The results taking into account the foregoing changes are indicated in table 2 For 1962 the negative relation is evident between the industry-scaled trade balances and the physical capital/labor ratio The skill ratio apparently provides a somewhat better fit than the human capital/labor ratio when entered separately, but there is evidence of collinearity when both ratios are included However, for 1969, the H/L variable performed better than the skill ratio which was based on 1961 data A test was conducted to determine if the coefficients had changed significantly between 1962 and 1969 As would be surmised, the only change of consequence was in the coefficient for H/L, which was significant at the 5 percent level[7] It is apparent from the values of the R^2 that we have accounted for only a relatively small proportion of the variation in the industry-scaled trade balances[8] The results nevertheless show that West Germany is implicitly exporting relatively more of the services of human capital and importing the services of physical capital

---

[5] Issues in the aggregation of physical and human capital are discussed in Stern (1975a)

[6] Lacking data for 1969, gross values of production for 1967 were used to scale the 1969 industry balance

[7] This test was based on an application of a test for general linear constraints for the coefficient vector given in Theil (1971, pp 143–44) and adapted for use here with the assistance of E Philip Howrey Details are available upon request

[8] As mentioned already, it would be desirable to include other variables, but data limitations preclude our doing so Our results may thus be subject to specification error unless the omitted explanatory variables are uncorrelated with the included explanatory variables, which is an assumption that cannot be readily assessed in the present context For further details, see, for example, Kmenta (1971, pp 392–95)

TABLE 2

ESTIMATES OF REGRESSION EQUATIONS EXPLAINING RATIO OF WEST GERMAN NET EXPORTS TO VALUE OF GROSS PRODUCTION FOR 1962 AND 1969

EQUATION	CONSTANT TERM	INDEPENDENT VARIABLES			N	R^2	F
		$\frac{K}{L}$	$\frac{H}{L}$	$\frac{E+T}{L}$			
1962a	020 (374)	– 002** (2 240)	.006* (1 836)		25	204	2 818*
1962b	026 (571)	– 002** (2 380)		021** (2 467)	25	281	4 297**
1962c	– 005 (102)	– 003** (2 650)	003 (1 132)	017* (1 915)	25	322	3 329**
1969a	– 036 (453)	– 002** (2 366)	.052** (2,430)		25	253	3 730**
1969b	015 (214)	– 002* (1 786)		025* (2 008)	25	200	2 742*
1969c	– 075 (934)	– 003** (2 648)	043* (2 027)	019 (1 558)	25	331	3 456**

NOTE — See table 1 footnotes

Since data were available on West German trade balances by industry vis-à-vis the other industrial countries and the LDCs, we next sought to determine if this distinction had any effect on the foregoing relationships The results are indicated in table 3 [9] For 1962 it is evident that the coefficient for the K/L ratio was negative for industry-scaled trade balances for both regions The skill ratio performed somewhat better than the H/L variable, and there were indications again of collinearity For 1969 the results for the K/L and H/L variables for industrial countries were similar to the results obtained for the total trade balances noted in table 2 For the LDCs only the H/L variable held up in the regressions reported In any event, it would appear from these results that West Germany is implicitly exporting relatively more of the services of human capital to both the industrial countries and the LDCs It is implicitly importing the services of physical capital from the industrial countries and only very weakly so with respect to the LDCs

The sample of 25 industries included seven that were natural-resource intensive or dependent to a large extent on nontraded goods [10] In order to determine how inclusion of these seven industries might affect the

---

[9] In principle, the coefficients for the industrial countries and LDCs in table 3 should sum to those shown in table 2, except that West German trade with the socialist countries has been excluded from the analysis If the regional dichotomy in table 3 had been comprehensive, it would have been desirable to impose adding-up constraints and use generalized least squares in order to improve the efficiency of estimation

[10] These were mineral oil, pulp, paper, and paperboard, printing and related industries, nonferrous metals, stone and earthern goods, food, beverages, and tobacco, and sawmills and woodworking It may be noted that the mining industry had already been excluded from the 25-industry sample

TABLE 3

ESTIMATES OF REGRESSION EQUATIONS EXPLAINING RATIO OF WEST GERMAN NET EXPORTS TO VALUE OF GROSS PRODUCTION FOR INDUSTRIAL COUNTRIES (INDC) AND LESS DEVELOPED COUNTRIES (LDC), 1962 AND 1969

		INDEPENDENT VARIABLES					
EQUATION	CONSTANT TERM	$\frac{K}{L}$	$\frac{H}{L}$	$\frac{E+T}{L}$	N	R^2	F
1962							
INDC(a)	022 (574)	– 002* (2 066)	004 (1 593)		25	174	2 317
INDC(b)	022 (687)	– 001** (2 295)		014** (2 357)	25	264	3 955**
INDC(c)	004 (112)	– 002** (2 452)	002 (902)	012* (1 870)	25	292	2 886*
LDC(a)	004 (264)	– 001** (2 226)	001* (1 810)		25	201	2 768*
LDC(b)	008 (634)	– 001** (2 130)		004* (1 948)	25	217	3 052*
LDC(c)	– 001 (295)	– 001** (2 470)	001 (1 226)	003 (1 402)	25	270	2 582*
1969							
INDC(a)	– 010 (175)	– 002** (2 647)	037** (2 437)		25	277	4 213**
INDC(b)	030 (582)	– 001* (2 031)		017* (1 879)	25	209	2 901*
INDC (c)	– 036 (0 618)	– 002*** (2 891)	032* (2 048)	013 (1 422)	25	340	3 613**
LDC(a)	– 020 (946)	– 0004 (1 4652)	012** (2 123)		25	174	2 323
LDC(b)	– 008 (410)	– 0002 (9027)		006* (1 744)	25	126	1 586
LDC (c)	– 029 (1 328)	– 0004 (1 6699)	010* (1 751)	004 (1 317)	25	237	2 179

NOTE –See table 1 footnotes

results, the regressions in tables 2 and 3 were reestimated A dummy variable was introduced separately for natural resource industries (0) and other industries (1), and multiplicative interaction terms were also formed with each of the independent variables in the model A heteroskedastic correction was applied to the OLS estimates to take into account the presumed variation between observations with regard to the two-industry groups [11] The results are shown in table 4

---

[11] The rationale for this procedure can be seen if we posit regression equations as follows

$$Y_i = \alpha_0 + \alpha_1 X_i + \varepsilon_i \qquad Z = 0, \tag{3}$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad Z = 1 \tag{4}$$

for the two-industry groups, where Y is the dependent variable, X refers to the independent variables, and the ε's are disturbances If we enter the dummy variable separately and interacted with the independent variables, the regression equation to be estimated is

$$Y_i = \gamma_0 + \gamma_1 Z_i + \gamma_2 X Z_i + \gamma_3 X_i + \varepsilon_i \tag{5}$$

For 1962 it is evident, especially for the LDCs, that the two groups of industries can be distinguished in terms of the coefficient characteristics of K/L and $(E + T)/L$. These results are much more clearly evidenced for 1969 In contrast to the results in tables 2 and 3, the H/L variable was not statistically significant It thus appears that the skill ratio may capture the relevant human capital characteristics more effectively than the wage-differential measure when the industry groups are separately identified We had hypothesized that the skill ratio might reflect industry differences in R & D, but this cannot be ascertained precisely without additional data

Regression coefficients for the levels of the relevant variables for the two industry groups are recorded in table 5 [12] While many of the individual coefficients were not statistically significant, it is evident that the natural-resource industries had positive coefficients for the K/L ratio and negative coefficients for the skill ratio, whereas the opposite pattern was recorded for the other manufacturing industries It can be concluded therefore from tables 4 and 5 that the factor services implicit in West German net exports differ depending on whether we are considering natural resource or other manufacturing industries This is shown to be the case for both total net exports and for net exports vis-à-vis the industrial countries and the LDCs [13]

One final thing that we did was to investigate the influence of nominal

---

Because the natural resource and other industries may be structurally different, the variance of the stochastic disturbances in eqq (3) and (4) may be presumed nonconstant This implies that OLS estimation of eq (5) will not be efficient and that a correction for heteroskedasticity is required We thus proceeded by first estimating eq (5) by means of OLS and then, on the basis of the residuals, calculated the standard error of estimate for each of the two-industry groups We then transformed the variables and constant terms by dividing them by each industry group's estimated standard deviation and ran OLS on the transformed eq (5) Details on the foregoing procedure, which is asymptotically equivalent to maximum likelihood estimation, can be found in Kmenta (1971, pp 264–66)

[12] Eq (3) above for the natural resource industries ($Z = 0$) can be written in terms of eq (5) as $Y_i = \gamma_0 + \gamma_3 X_i + \varepsilon_i$, from which it follows that $\alpha_0 \equiv \gamma_0$ and $\alpha_1 \equiv \gamma_3$ Similarly, for other industries ($Z = 1$), we have $Y_i = (\gamma_0 + \gamma_1) + (\gamma_2 + \gamma_3)X_i + \varepsilon_i$ so that $\beta_0 \equiv \gamma_0 + \gamma_1$ and $\beta_1 \equiv \gamma_2 + \gamma_3$ and $\gamma_1 \equiv \beta_0 - \alpha_0$ and $\gamma_2 \equiv \beta_1 - \alpha_1$ Thus, the regression coefficients for the natural resource industries are the ones shown under K/L, H/L, and $(E + T)/L$ in table 4, whereas the regression coefficients for the other industries are obtained by summing the aforementioned ones with the respective dummy-variable interactions noted in table 4 It should be noted that the coefficients recorded in table 5 have been calculated with more decimal places than indicated in table 4

[13] See Baldwin (1971, pp 139–41) for an analysis of the determinants of U S net exports by region for 1962 While our results cannot be compared directly with his because of differences in measurement of variables and our use of a less detailed breakdown, there is nevertheless a broad similarity in the performance of our skill ratio for the two types of regions Similarly, while our results cannot be compared directly with those obtained by Fels (1972, 1973), our conclusions are somewhat different from his in suggesting that the services of human capital in West German net exports are important with respect to both the industrial countries and the LDCs

TABLE 4

Estimates of Regression Equations Explaining Ratio of West German Net Exports to Value of Gross Production for 1962 and 1969, with Dummy Variables Distinguishing Natural Resource (0) and Other Industries (1)

Equation	Constant Term	$\frac{K}{L}$	$\frac{H}{L}$	$\frac{E+T}{L}$	Dum	(Dum) $\left(\frac{K}{L}\right)$	(Dum) $\left(\frac{H}{L}\right)$	(Dum) $\left(\frac{E+T}{L}\right)$	N	R^2	F
		Independent Variables									
1962											
All countries	– 190*	003	005	– 046	240*	– 004	– 004	060*	25	616	4 082***
	(1 800)	(1 093)	(472)	(1 526)	(2 036)	(1 191)	(329)	(1 929)			
INDC	– 113	001	001	– 016	164*	– 002	– 001	024	25	537	2 958**
	(1 493)	(694)	(190)	(744)	(1 882)	(930)	(066)	(1 081)			
LDC	– 039	001**	001	– 024***	040	– 001*	– 001	029***	25	747	7 938***
	(1 464)	(2 268)	(412)	(3 193)	(1 424)	(1 998)	(300)	(3 695)			
1969											
All countries	– 275	004	022	– 091**	285	– 006*	010	111**	25	676	5 191***
	(1 508)	(1 223)	(232)	(2 265)	(1 466)	(1 765)	(111)	(2 688)			
INDC	– 139	003	– 007	– 055*	171	– 005*	036	066**	25	627	4 234***
	(1 024)	(1 155)	(105)	(1 839)	(1 167)	(1 886)	(500)	(2 149)			
LDC	– 088**	001*	014	– 034***	698	– 001*	– 009	041***	25	760	7 391***
	(2 116)	(1 893)	(635)	(3 757)	(1 600)	(1 852)	(404)	(4 378)			

Note – See table 1 footnotes

TABLE 5

REGRESSION COEFFICIENTS FOR WEST GERMAN
NATURAL RESOURCE AND OTHER INDUSTRIES, 1962 AND 1969

	NATURAL RESOURCE INDUSTRIES			OTHER INDUSTRIES		
EQUATION	$\frac{K}{L}$	$\frac{H}{L}$	$\frac{E+T}{L}$	$\frac{K}{L}$	$\frac{H}{L}$	$\frac{E+T}{L}$
1962						
All countries	003	005	– 046	– 001	001	014*
	(1 093)	(472)	(1 526)	(593)	(419)	(1 776)
INDC	001	001	– 016	– 001	001	008
	(694)	(190)	(744)	(629)	(342)	(1 265)
LDC	001**	001	– 024***	– 0001	0003	004***
	(2 268)	(412)	(3 193)	(.1907)	(5309)	(3 215)
1969						
All countries	004	022	– 091**	– 003	032*	020**
	(1 223)	(232)	(2 265)	(1 361)	(1 797)	(2 147)
INDC	003	– 007	– 055*	– 003	028*	011
	(1 155)	(105)	(1 839)	(1 632)	(1 927)	(1 494)
LDC	001*	014	– 034***	– 0002	005	006***
	(1 893)	(635)	(3 757)	(4036)	(1 375)	(3 584)

NOTE —See table 1 footnotes

tariffs and effective tariffs on West German net exports The variables had the wrong sign, however, and therefore the results are not reported in detail here [14]

III. Conclusion

While our study has been circumscribed due to data limitations, we have nonetheless established that West Germany is implicitly exporting the services of human capital and implicitly importing the services of physical capital for other than natural resource industries This result appeared for net exports with respect to both industrial countries and LDCs The opposite pattern was observed for the natural-resource industries The conclusion to be drawn therefore is that the distinguishing characteristics of natural resource and other industries must be given explicit recognition in studying the determinants of the structure of trade

---

[14] Tariff data were available from Fels (1972) for 1970 for the 18 manufacturing industries, exclusive of the natural resource industries Assuming that relatively higher tariff rates had a restrictive effect on imports, the sign of the tariff variable was expected to be positive since industries with higher tariffs would have smaller net import ratios We thus entered the nominal and effective tariff rates, with West German intra-EEC trade eliminated, into our regressions Both tariff variables had negative coefficients in all the regressions, but only the effective tariff coefficients were statistically significant. These results thus suggest a different direction of causation than hypothesized so that what we really need is a model that can account for both the imposition and consequences of tariffs Such an inquiry must be left for future investigation

We have also shown that the scaling of variables affected the results substantially; the coefficients on physical and human capital were different in sign and magnitude, making the aggregation of the two types of capital questionable in studies of the factor services embodied in trade; the coefficient on the human capital variable appeared to have increased markedly between 1962 and 1969; and the influence of tariffs was the opposite of what was expected, suggesting that we need to consider how net imports determine tariffs as well as vice versa

We have by no means provided a comprehensive explanation of the factors reflected in West German trade In particular, it would be desirable to investigate the importance of expenditures on R & D, economies of scale, and innovative performance Furthermore, it would be interesting to make relative comparisons among countries if the requisite information could be obtained

References

Baldwin, R E "Determinants of the Commodity Structure of U S Trade" *A E R* 61, no 1 (March 1971) 126–46

Branson, W H "U S Comparative Advantage Some Further Results" *Brookings Papers Econ Activity* 3 (1971) 754–59

——— "Factor Inputs, U S Trade, and the Heckscher-Ohlin Model" Institute for International Economic Studies, Seminar Paper no 27, Stockholm, May 1973

Branson, W H, and Junz, H B "Trends in U S Trade and Comparative Advantage" *Brookings Papers Econ Activity* 2 (1971) 285–338

Fels, G "The Choice of Industry Mix in the Division of Labour between Developed and Developing Countries" *Weltwirtschaftliches Archiv* 108, Heft 1 (1972) 71–121

——— "The Export Needs of Developing Countries and the Adjustment Process in Industrial Countries" In *International Division of Labour Problems and Perspectives*, edited by Herbert Giersch International Symposium Kiel Institut fur Weltwirtschaft, 1973

Hufbauer, G C "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods" In *The Technology Factor in International Trade*, edited by R Vernon New York· Columbia Univ Press, 1970

Katrak, H "Human Skills, R and D and Scale Economies in the Exports of the United States and United Kingdom" *Oxford Econ Papers* 25, no 3 (November 1973) 337–60

Kmenta, J *Elements of Econometrics* New York Macmillan, 1971

Leamer, E E "The Commodity Composition of International Trade in Manufactures An Empirical Analysis" *Oxford Econ Papers* 26, no. 3 (November 1974) 350–74

Morrall, J F., III *Human Capital, Technology, and the Role of the United States in International Trade* Gainesville Univ Florida Press, 1972

Roskamp, K W, and McMeekin, G. C. "Factor Proportions, Human Capital and Foreign Trade The Case of West Germany Revisited" *Q J E* 82, no. 1 (February 1968) 152–60.

Stern, R. M. "Capital-Skill Complementarity and U.S. Trade in Manufactures." In *Quantitative Studies of International Economic Relations,* edited by H. Glejser. Amsterdam: North-Holland, 1975 (*a*)

——— "Testing Trade Theories." In *International Trade and Finance: Frontiers for Research,* edited by P B. Kenen New York. Cambridge Univ Press, 1975. (*b*)

Theil, H. *Principles of Econometrics.* New York· Wiley, 1971

The Property Rights Theory of the Firm: Empirical Results from a Natural Experiment

H. E. Frech III

University of California, Santa Barbara

In the property rights theory of the firm (Williamson 1969, 1970; Alchian 1961, Alchian and Kessel 1962, Alchian and Demsetz 1972, Furubotn and Pejovich 1972) a connection is made between the property rights in the firm held by the decision maker and the behavior of that decision maker This paper compares the performance of firms with different types of property rights in providing a standardized product under contract to the federal government[1]

The standardized product is physician services insurance (Part B) claim processing for the Medicare program Both private property firms (commercial profit-seeking insurance companies) and attenuated property rights firms (nonprofit mutual and Blue Shield firms) provide these services to the Medicare administration for different administrative areas Thus, the Social Security Administration has provided an opportunity for observing the influence of alternative specifications of property rights on the performance of the firms

Theory

Under private property rights the decision maker has the right to (1) decide about the use of the firm's resources, (2) keep the residual (total revenue minus contractural costs), and (3) capitalize any wealth gains of the firm by selling his rights

---

An earlier version of this paper was presented at the Western Economic Association meetings, Las Vegas, June 1974 Thanks are due to the participants of that session, the editor, Paul B Ginsburg, Kenneth W Clarkson, Michael Ward, Sam Peltzman, and Jack Hirshleifer for helpful comments Robert Topel provided valuable research assistance and excellent comments Ronald J Vogel gave vital assistance in obtaining data and illuminating criticism Special thanks are due to David M Barton, who shared in the early conception and planning of the research If he had been able to continue, he would be listed as coauthor Finally, I am grateful to the University of California, Santa Barbara Academic Senate Committee on Research for financial support

[1] For discussion of the nature of the contract see Blair, Ginsburg, and Vogel (1975)

[*Journal of Political Economy*, 1976, vol 84, no 1]

© 1976 by The University of Chicago All rights reserved

As property rights are attenuated for a nonprofit firm, the decision maker cannot keep the residual Further, he cannot capitalize additions to the firm's wealth by selling the rights to future income flows

The attenuated rights lead to different behavior than private property rights in two ways First, the attenuated rights effectively reduce the price of nonpecuniary amenities (pleasant offices and colleagues, short working hours, greater output) to the decision makers This leads to choice of lower firm wealth and more nonpecuniary benefits

Using the approach of Williamson (1970), the effects of attenuated property rights can be formalized The decision maker maximizes a strictly concave utility function over wealth, W, and nonpecuniary benefits, N, $U(W, N)$, subject to the concave technological and market constraint on the possibilities combinations of W and N, $T(W, N) \leq 0$, and the non-private-property constraint[2] limiting wealth to the present value of salary, S, $W - S \leq 0$ Ignoring corner solutions, the marginal condition is $(dW/dN) = -(\partial U/\partial N)/[(\partial U/\partial W) - \lambda_2]$, where λ_2 is the Legrangion multiplier for the non-private-property constraint

The interpretation of this condition is straightforward Under private-property rights, $\lambda_2 = 0$, and the technical rate of exchange W and N is equated to the rate of substitution in utility terms This corresponds to point *Op* in figure 1

However, under attenuated property rights, $\lambda_2 > 0$ Thus (dW/dN) is lower (more negative) This can only occur where W is lower and N is higher, corresponding to *Os* is figure 1

Thus, attenuated property rights lead to a choice of lower firm wealth and greater nonpecuniary benefits If these nonpecuniary benefits include elements other than quantity and quality of output, the nonprofit firm will exhibit higher production costs, for the same output, than the profit-seeking firm

The second way in which behavior altered is that the attenuated rights reduce the gains to the takeover of poorly managed firms The gains of efficient new management are of less value to the new owner because he could neither retain the residual in wealth terms nor sell his ownership after improving the management of the firm

The result of these reinforcing effects is that, in Alchian and Demsetz's words (1972, p 790), "One should, therefore, find greater (management) shirking in nonprofit enterprises "

An aside on the survival of nonprofit firms —The analysis presented above implies that nonprofit firms will operate less efficiently than profit-seeking firms If entry were free, competition would lead to the demise of nonprofit

[2] It may be possible to increase wealth by giving up some other benefits, but this only complicates the analysis

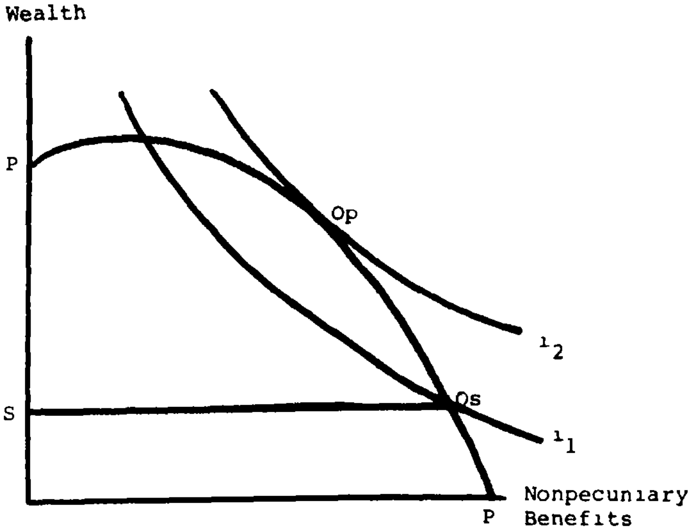


FIG 1 —Property rights and the optimal combination of firm wealth and managerial nonpecuniary benefits

firms There are two reasons why this predicted disappearance of nonprofit firms does not occur

First, and simplest, government regulation may cause cost disadvantages for profit-seeking firms This will be discussed below

Second, the nonprofit legal form has important advantages for a firm whose output is subsidized by charitable contributions Donors (either private parties or the government) do not want the firm's owners' wealth maximized—they seek to raise the firm's output By preventing appropriation of the residual, they can be assured that donations and grants must be spent within the firm If output is an important nonpecuniary benefit to the firm decision maker, donations will increase output

Past Studies

In spite of the intuitive appeal and apparent importance of the analysis, there have been relatively few attempts to measure empirically the impact of alternative property rights specifications on costs of operation The ones which have been undertaken (Nichols 1967, Clarkson 1971, 1972, Davies 1971, Spiller 1972) suffer from the problem that the output is not quite held constant nor controlled for in the analysis. The result is some degree of ambiguity concerning the independent influence of property rights structure versus the difference in costs of supplying different products. Another important weakness in much of the work is a lack of statistical testing (Clarkson's work is an important exception). This study examines the data on Medicare insurance claim processing by intermediary firms of differing property rights structures This provides a natural experiment, where the data allow one to attribute with some confidence differences to property rights structures, since the specification of output and the legal environment are determined by contract with the Social Security Administration [3]

Industry Structure

The health insurance industry, from which Medicare intermediaries are chosen by the Social Security Administration, is largely made up of three types of firms, ordinary profit-seeking insurers, mutual insurers, and Blue Cross and Blue Shield nonprofit firms which provide physician and hospital insurance, respectively

By the argument of the "Theory" section above, the nonprofit insurers should be unable to compete with the more efficient profit-seeking insurers Therefore, the survival of the mutual and Blue Cross–Blue Shield firms deserves some comment

In the case of the mutual firms, it appears that their form allows them to undercut the regulated insurance price by paying dividends to insurers (Kimball 1960) Further, as is discussed below, their actual property rights may not be very different from profit seeking firms

Blue Cross and Blue Shield firms survive because of quite strong regulatory and tax advantages in almost every state (Frech 1974) In many states, Blue Cross and Shield firms pay no taxes on property or premium income, while profit-seeking and mutual firms must pay such taxes Premium taxes alone average about 2 percent of gross premium volume In the majority of states, special laws have been passed granting tax and regulatory advantages to these firms

The Experiment

The Social Security Administration contracts with private health insurance firms for claim processing services These firms deal directly with medical care providers and consumers A government report on the Medicare program states that "the carriers' principal function is to determine whether charges are allowable (reasonable) and to make

---

[3] For more detail on the nature of the relationship of the Social Security Administration to the independent intermediaries, see Blair, Ginsburg and Vogel (1975) and U.S. Department of Health, Education, and Welfare (1973*b*).

TABLE 1

MEDICARE PROCESSING PERFORMANCE AND OWNERSHIP

	Cost per Dollar Processed	Average Processing Time (Days)	Errors per $1,000 Processed
Profit-seeking firms ($N = 12$)	$0 0715 (0 0178)	18.70 (10 34)	0 1844 (0 0717)
Nonprofit firms ($N = 66$)	0 1039 (0 0331)	33 59 (19 83)	0 4426 (0 7576)
Approximate t-value for difference of means	4 908	3.565	2 706
Degrees of freedom	29 9	31.3	70 9
Statistical significance (one-tail test)	< 0005	0009	.005

NOTE —If variances are assumed to be different and unknown, the appropriate test statistic is an approximate t value (Yamane 1967, pp 522–24) Standard deviations are shown in parentheses

payment" (U S Department of Health, Education, and Welfare 1973*b*). Thus, the carriers have no role in such common insurance activities as classification of risks and determination of benefit packages The narrowness of the service and its contractural (rather than market) determination are important advantages Unmeasured variation in type of service is likely to be very low

As is obvious from the large variation in the data on costs and performance, firms and their geographical regions were not selected on the basis of competitive bidding Further, no carrier has been dropped for poor performance Since the selection and retention of intermediaries was not based on cost minimization, efficiency cannot be inferred from survivorship

The Empirical Analysis

Property rights and performance —The first step to determine the influence of property rights structure on performance is to examine the actual performance of the firms The data presented in table 1 give the means, standard deviations, and statistical significance level of the difference in means for average cost per dollar processed, average processing time, and errors per dollar processed for nonprofit and profit-seeking Medicare Part B intermediaries, based on data from the Social Security Administration (U S Department of Health, Education, and Welfare 1973*a*) [4] Nonprofit firms include Blue Shield, one government agency, one prepaid group practice, and mutual (owned by policyholders) firms[5] As one can see, the superiority of the private property firms is striking. The probability of these performance differences occurring by chance is apparently very low

---

[4] A nonprofit firm with very poor performance was deleted from the sample because of incomplete data on errors

Scale and performance—Observation of the scale of operation of the intermediary operations of the profit-seeking and nonprofit firms reveals that the average nonprofit firm operation is smaller ($6 679 million contrasted to $8 914 million per month) It may be interesting to examine how much of the lower performance of the nonprofit firms can be associated with the lower scale of operation Further, the nonprofit firms produce at a lower scale in the dimension of what Hirshleifer (1962) calls the breadth of production The size of the average claim of the nonprofit firm is lower than that of the profit firm ($32 12 vs $45) Following Hirshleifer's reasoning, aggregating the same dollar value of payments into fewer larger claims would lower costs Thus, I will examine the relationship of performance to these two dimensions of scale to see if the poorer performance of the nonprofit firms can be explained entirely by low levels in these two scale dimensions

For purposes of the estimated functions, the scale variables must be taken as exogenous due to lack of data Based on the same data as the comparison of means, functions have been estimated relating performance to the two dimensions of scale and the property right structure of the firm

Estimated equations—The simplest version of the cost function takes the following form (in logs) $\mathrm{CPAY} = \mathrm{C} + \beta_1\ \mathrm{PAY} + \beta_2\ \mathrm{CLSZ} + \beta_3\ \mathrm{NP}$, where C is a constant, CPAY is the cost of processing per dollar payment[6] made by the intermediary PAY is a scale variable, the dollar value of the payments made by the intermediary, CLSZ is the size of the average claim, a measure of the breadth of production, and NP is a dummy variable (not in logs) which takes on the value zero for a private property and one for a nonprofit firm[7]

For the other output dimensions, average processing time (in days), TIME, and the number of errors per dollar processed, EAPY, and the reduced-form equation is of the same form. Below are the equations, which were estimated using ordinary least squares (t-statistics in parentheses):

---

[5] Although mutual insurers are nominally nonprofit firms, their actual constraints may be quite similar to profit-seeking firms First, the remuneration of the top-management group is not closely regulated Second, mutual firms often have close corporate ties with profit-seeking firms Thus, one might expect their behavior to be more like profit-seeking firms When mutuals are compared with other nonprofit firms they perform significantly better in terms of cost and processing time, but slightly (and statistically insignificantly) worse in the dimension of accuracy Apparently their property rights are somewhat less attenuated than those of the typical nonprofit firms

[6] A reasonable alternative definition of average cost is cost per claim processed This amounts to deflating cost and errors by the number of claims, rather than by dollar volume Equations (1)-(3) were reestimated, using this definition The results were virtually identical, except for slightly worse fit to the data

[7] More complex functional forms, such as quadratic and dummied slopes, gave similar results, but with notably lower t-values and R^2 (corrected for degrees of freedom)

$$\begin{array}{lrrrrr}
\mathrm{CPAY} = & -0.703 & -0.051\ \mathrm{PAY} & -0.390\ \mathrm{CLSZ} & +0.152\ \mathrm{NP} & \\
 & & (-2.598) & (-3.517) & (2.219) & (1) \\
 & & \multicolumn{2}{c}{N = 78,\ R^2 = .605} & & \\
\mathrm{TIME} = & 1.931 & -0.080\ \mathrm{PAY} & +0.436\ \mathrm{CLSZ} & +0.600\ \mathrm{NP} & \\
 & & (-1.482) & (1.435) & (3.179) & (2) \\
 & & \multicolumn{2}{c}{N = 78,\ R^2 = .140} & & \\
\mathrm{EPAY} = & 5.815 & +0.032\ \mathrm{PAY} & -2.084\ \mathrm{CLSZ} & -0.550\ \mathrm{NP} & \\
 & & (0.333) & (-3.723) & (-1.618) & (3) \\
 & & \multicolumn{2}{c}{N = 78,\ R^2 = .294} & &
\end{array}$$

The cost function is most interesting. This simple model, with only two dimensions of scale and the property rights variable, explains over 60 percent of the cost variation. The effect of a nonprofit property rights structure is to raise costs by about 15 percent. And the difference is significant at the 5 percent level.

The equation for the determination of TIME does not fit as well, explaining only about 14 percent of variation of TIME. However, the results are quite interesting and reasonable. Scale has a negative effect on processing time, while the size of *each claim* (reflecting claim complexity, in part) increases TIME. Note that the nonprofit firms perform substantially worse in this dimension, NP increases TIME by about 60 percent, with the coefficient significant at the 0.1 percent level.

The equation for errors per dollar of payment fits better than the one for TIME, explaining about 30 percent of the variation in EPAY. Scale is essentially unrelated to EPAY, while the average claim size is strongly negatively related, presumably because processing the same number of dollars with fewer claims leads to fewer opportunities for errors. Turning to the property rights variable, EPAY is the only output dimension for which controlling for the two scale dimensions reverses the result that nonprofit firms are less efficient. However, the NP variable is not significant at the 5 percent level (the t-statistic of -1.618 is significant at approximately the 10 percent level).

Unfortunately, unless one rejects the relationship of NP to EPAY as statistically insignificant, the results are ambiguous to the issue of whether the choice of scale is the main determinant of nonprofit inefficiency or whether the entire cost surface is higher for such firms. There are two possible ways to get some indication concerning whether private property rights lowers the firm's cost surface. The most natural is simply to enter the variable EPAY as an explanatory variable in equations (1) and (2) to see if nonprofit firms have higher costs and slower processing times holding errors constant. Doing that results in the following equations.

$$\begin{array}{llll} \text{CPAY} = -0.745 & -0.051\ \text{PAY} & -0.375\ \text{CLSZ} & \\ & (-2.592) & (3.075) & \\ & +0.156\ \text{NP} & +0.007\ \text{EPAY} & (4) \\ & (2.224) & (0.307) & \\ & N = 78,\ R^2 = .605 & & \end{array}$$

$$\begin{array}{llll} \text{TIME} = 1.395 & -0.083\ \text{PAY} & +0.628\ \text{CLSZ} & \\ & (-1.547) & (1.905) & \\ & +0.648\ \text{NP} & +0.921\ \text{EPAY} & (5) \\ & (3.413) & (1.444) & \\ & N = 78,\ R^2 = .164 & & \end{array}$$

Taken at face value, these equations indicate that even holding EPAY constant private property firms are more efficient in terms of costs and processing time. Inclusion of EPAY in the two equations makes virtually no difference. This would seem to add confidence to our results.

Unfortunately, there is a serious econometric problem with this approach. EPAY is not exogenous. In fact, the errors in equation (3) are doubtlessly positively correlated with those of equations (4) and (5). If the errors in those equations represent firms which are either poorly managed or operate in a high-cost environment, these firms are likely to perform badly in all three output dimensions. Thus, the errors will be positively correlated and the coefficient of EPAY in equations (4) and (5) will be biased upward. Since the true tradeoff between TIME or CPAY and EPAY is clearly negative (errors can only be reduced by processing slower or more carefully at a higher cost), the positive coefficients of equations (4) and (5) for EPAY indicate the extent of the simultaneous-equations bias and do not estimate the shape of the production possibility tradeoff among the three dimensions of output.

Since we are not specifically interested in the shape of this tradeoff, the incorrect sign on EPAY creates no problem. However, the bias indicates that the equations (4) and (5) do not control for EPAY, and therefore one cannot reject the hypothesis that the nonprofit firms are slower and have higher costs solely because they are more careful, which results in fewer errors (holding scale constant).

An alternative approach is to take the point estimates of the reduced-form equations as accurate and see how high a price one would have to place on the fewer errors of the nonprofit firms in order for their total costs to be equal to those of the private property firms (here we ignore the advantages of quick processing by the profit-seeking firms).

Working through the calculations for this approach I find that the implicit cost of an accounting error would have to be implausibly high at about $67.88 for the nonprofit performance to be equal to the profit-seeking firms [8]

Summary and conclusion.—The evidence suggests that private property firms are more efficient as Medicare processing intermediaries than nonprofit firms. Nonprofit firms incur processing costs per dollar processed which are 45 percent higher Their processing takes 80 percent longer Further, nonprofit firms make 140 percent more errors per dollar processed These differences are significant at very high levels of confidence

Exploration by regression analysis of the impact of economies of scale in overall operations and in aggregating claims showed that some, but not all of this difference was due to inefficiently low scale in both dimensions for nonprofit firms Apparently nonprofit firms not only chose inefficient points on their cost surface, but the entire surface is higher for them

References

Alchian, A "Some Economics of Property " RAND Corporation, P-2316 Santa Monica, Calif RAND Corp , May 16, 1961

——— "Corporate Management and Property Rights " In *Economic Policy and the Regulation of Corporate Securities*, edited by H Manne Washington · American Enterprise Inst Public Policy Res , 1969

Alchian, A , and Demsetz, H "Production, Information Costs and Economic Organization " *A E R* 62 (December 1972) 777–95

Alchian, A , and Kessel, R "Competition, Monopoly and the Pursuit of Money." In National Bureau of Economic Research, *Aspects of Labor Economics* Princeton, N J Princeton Univ Press, 1962

Blair, R D , Ginsburg, P B., and Vogel, R J "Blue Cross–Blue Shield Administrative Costs A Study of Non-Profit Health Insurers." *Econ Inquiry* 13 (June 1975) 237–51

Clarkson, K *Property Rights, Institutional Constraints, and Individual Behavior. An Application to Short Term General Hospitals* Ph D dissertation, Univ California, Los Angeles, 1971

——— "Some Implications of Property Rights in Hospital Management " *J Law and Econ* 15 (October 1972) 363–84

Davies, D G "The Efficiency of Public versus Private Firms The Case of Australia's Two Airlines " *J Law and Econ* 14 (April 1971) 149–66.

Frech, H E , III *The Regulation of Health Insurance* Ph D dissertation, Univ California, Los Angeles, 1974

---

[8] We seek the implicit price of an accounting error, Pe, which would nullify the predicted difference in CPAY between the two types of firms That is, the solution to ΔCPAY = Pe ΔEPAY, where ΔCPAY is the predicted difference between the two types of firms in CPAY in equation (1) and ΔEPAY is the predicted difference in EPAY from equation (2) expressed in terms of errors per dollar processed (rather than per $1,000 processed as in table 1 and equation [3]) Using the estimated coefficients on NP from the equations and the means of sample we solve

$$(0\ 0989)(0\ 152) = Pe(0\ 403)(10^{-3})(0.550),$$
$$Pe = \$67\ 88$$

Furubotn, E., and Pejovich, S. "Property Rights and Economic Theory: A Survey of Recent Literature." *J. Econ. Literature* 10 (December 1972): 137–62.

Hirshleifer, J. "The Firm's Cost Function: A Successful Reconstruction?" *J. Bus.* 35 (July 1962): 235–55.

Kimball, S. L. *Insurance and Public Policy: A Study in the Legal Implementation of Social and Economic Public Policy, Based on Wisconsin Records 1835–1959.* Madison: Univ. Wisconsin Press, 1960.

Nicols, A. "Stock versus Mutual Savings and Loan Associations: Some Evidence of Differences in Behavior." *A.E.R.* 57 (May 1967): 337–46.

Spiller, R. "Ownership and Performance: Stock and Mutual Life Insurance Companies." *J. Risk and Insurance* 34 (March 1972): 17–25.

U.S. Department of Health, Education, and Welfare, Social Security Administration, Bureau of Health Insurance. *Memorandum: SMI Carrier Operations: Selected Workload and Cost Data, January–March 1970.* Washington: Dept. Health, Education, and Welfare, July 7, 1973 (*a*).

U.S. Department of Health, Education, and Welfare, Social Security Administration, Office of Research and Statistics. *Medicare: Health Insurance for the Aged, 1971, Section 2: Enrollment.* Washington: Government Printing Office, 1973 (*b*).

Williamson, O. E. "Corporate Control and the Theory of the Firm." In *Economic Policy and the Regulation of Corporate Securities,* edited by H. Manne. Washington: American Enterprise Inst. Public Policy Res., 1969.

———. *Corporate Control and Business Behavior: An Inquiry into the Effects of Organizational Form on Enterprise Behavior.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Yamane, T. *Statistics: An Introductory Analysis.* 2d ed. New York: Harper & Row, 1967.

Thornton's Estimates of Efficiency Losses in Soviet Industry: Some Fixed-Point-Method Recalculations

John Whalley

London School of Economics

A number of empirical estimates of static efficiency losses resulting from factor market interferences have appeared in recent years [1] What would appear to be the first set of estimates of efficiency losses arising from the pursuit of centrally directed allocation procedures have recently been presented for Soviet industry by Thornton (1971). Western economists have often suggested that the allocation procedures pursued in the Soviet economy result in an allocation of resources that is economically inefficient, and the allocation of capital goods between industries has been particularly stressed [2] It is this feature which Thornton examines numerically for the first time

Thornton's procedures involve the reconstruction of a set of financial accounts for Soviet industry in a form not published in the Soviet Union The financial accounts are assumed to be generated by a set of quantity allocations and commodity and factor prices These prices are assumed to have sufficient allocative significance that one can claim that value marginal products of factors in different uses are not equalized On the basis of these figures a simulation of the behavior of the economy in a regime where value marginal products of factors are equalized is then used to assess the inefficiency of the existing allocation The concern here is with Thornton's results in the context of efficiency loss estimates for market economies and, further, with the methods used for such a simulation

The interpretation of Thornton's exercise is, however, worth some clarification before moving on to examine her results The role and

---

I am grateful to Professors H G Johnson and J Thornton for their comments on an earlier draft and to the seminar in "Economics of the Communist World" at the London School of Economics for their helpful discussion of the material of this paper

[1] There is an extensive literature, particularly on the operation of tax systems (see, e g , Harberger 1966, Shoven and Whalley 1972, Whalley 1973), a number of other interferences such as trade union activity (Johnson and Mieskowski 1970) have also been examined along similar lines

[2] See the discussion in Bergson (1964, chap 14)

[*Journal of Political Economy* 1976, vol 84 no 1]
© 1976 by The University of Chicago All rights reserved

structure of the Soviet pricing system which underlies Thornton's accounts have been carefully studied in the West over a number of years The consensus would appear to be that, even though the Soviet economy is centrally planned in terms of quantities, the pricing system has a somewhat limited allocative significance. There is, however, a genuine area of doubt as to the degree to which the Soviet price structure is, in fact, "passive" in the sense of providing purely accounting prices In a limiting case, the underlying real allocation within the economy could be changed from an efficient to an inefficient allocation along with a change in the accounting prices such that the industry-wide financial accounts were unchanged To the extent that Soviet prices are "passive" in this way,[3] little quantitative inference in the way of economic efficiency can be confidently drawn from a simulation analysis of Soviet financial accounts The analysis in this note, as in Thornton's original paper, is predicted by an acceptance of a limited allocative role of the Soviet pricing structure, and it is this acceptance that allows the present equilibrium analysis of Soviet allocation procedures

In the light of efficiency loss estimates of factor market distortions in market economies, the estimates obtained by Thornton appear to be on the small side for the size of distortion considered (in the order of 3 percent of value added in Soviet industry) Thornton's calculations are made using the same general equilibrium structure as has been adopted for market economy loss estimates, but the distortion parameters involved considerably exceed those previously considered in work on market economies

The purpose of this note is to report some recalculations of Thornton's estimates which test their accuracy and sensitivity to various parameterizations of the model Thornton implicitly solves a numerical general equilibrium model but does so by using approximation procedures and under restrictive parametric assumptions Recent developments have made approximation procedures for the solution of equilibrium problems redundant, as full general equilibrium methods can be applied directly For some parameterizations of general equilibrium models it has been shown approximation procedures may be quite unreliable, for others less so

In addition to a concern over computational procedures, it is also the case that Thornton restricts attention to the case of all industries having Cobb-Douglas production functions, an assumption which is unsatisfactory in view of the finding by Weitzman (1970) of a low elasticity of substitution in Soviet industry The computational methods used here easily allow multi-industry efficiency loss calculations to be made for different values of the elasticities of substitution.[4]

---

[3] See the discussion of this issue presented in Wiles (1962)

[4] These methods are somewhat technical in detail, and the interested reader is referred to Shoven and Whalley (1972, 1973) and Scarf (1973) for further details

The assumptions on the Soviet pricing structure implicit in these calculations should, however, be reemphasized.[5] It is assumed that the labor market in the Soviet industrial sector is competitive and that an implicit capital-good market operates through charges levied by the central authority which are such as to clear the market were decentralization to be introduced It is often argued that Soviet labor markets are competitive to a tolerable approximation which, if reasonably accurate, leaves the reliability of the capital charges as possible capital market clearing decentralized prices subject to doubt, one must, of course, confess that a question mark (in perhaps heavy pencil) would have to be placed against this particular assumption.

Thornton's Estimates of Efficiency Losses in Soviet Industry

In an economy with many industries, each of which uses two factor inputs (capital and labor services) which may be substituted for each other, the conditions for an efficient allocation of productive factors require that the ratio of marginal products of factors be the same in all industries In a competitive market environment this will be assured if the same payment per unit is made to factors of production in all possible uses In market economies, factor market interferences have been characterized as resulting in factor rewards which vary according to industrial location As factors of production are assumed to "flow" between industries in such a way as to equalize the "net-of-distortion" returns, inequality in the gross returns across industries (the ratios of marginal products) results, yielding an inefficient allocation [6]

In the case of Soviet industry examined by Thornton it is primarily the application of investment criteria on a different basis for various sub industries (sectors) which it is suggested results in an inefficient allocation of productive resources In Soviet industry the application of payback-period criteria for investment appraisal typically involves different payback periods for each sector This may be represented as a situation in which differential effective charges are levied by the state by way of required returns to capital. In this way Thornton's model becomes analogous with those used in the analysis of factor market distortions in market economies. Those sectors with short payback periods tend to receive too little capital and those with long payback periods too much capital relative to that which they might receive if an efficient allocation were being sustained for the same output mix.

---

[5] One should also mention certain aspects of Thornton's calculations which are not examined in this note No changes are made in the model or the data used by Thornton as the focus is on the sensitivity and reliability of estimates There are undoubtedly refinements of this form which one might contemplate for these calculations. For instance, agriculture and services are not included, so the potential efficiency gains from movement of Soviet capital goods out of the industrial sector altogether are not captured No dynamic considerations are involved Equally, there is inevitably room for dispute over particular features of the data, and a reexamination of sources and interpretation of particular figures may affect the results

[6] Unless, by chance, the separate systems of distortions are offsetting, e g , industry A pays 20 percent tax on hirings of capital services and also 20 percent on hirings of labor services, so the ratio of marginal products (gross of tax returns) is the same as in industry B (where no tax operates)

Thornton's procedure for estimating efficiency losses arising in Soviet industry is first to determine the parameters of the model (the production functions and the differentials in factor charges by sector) and then make some assessment of the behavior of the economy in the absence of differential charges.[7] The production functions are assumed to be of Cobb-Douglas form, and the distortion parameters are estimated by examining data for components of Soviet industry from Thornton's reconstructed financial accounts (13 sectors within Soviet industry are considered). Data on the sector returns to capital[8] and the wage-bill[9] sector for 1960 and 1964 are used (the sum for each year being value added within each sector).

Capital stock data by sector are used to give an estimate of the capital charge per physical unit employed.[10] The differential of the imputed charge for each sector over the smallest charge calculated for all of the sectors may be thought of as the distortion parameter (one parameter for each sector).

In a market context, Harberger (1966) has been concerned with the efficiency loss from a tax differential in the United States of 85 percent on net-of-tax capital income between low taxes and highly taxed sectors. Thornton's distortions are much more substantial than this in terms of the largest distortion, although the graduation of distortions across industries is an offsetting feature.[11] Thornton's estimates of efficiency losses are in the order of 3 percent of value added in Soviet industry, which compare with Harberger's estimates of 1 0–1.5 percent of total product (for an economy-wide model).

---

[7] Thornton presents estimates both for 1960 and 1964, the figures for 1960 being given both for net and gross returns to capital in each industry (the difference being capital consumption by industry). Estimates using the net return to capital data are considered here. A justification for this is that a form of stationary state where the capital stock is being preserved is implicitly being considered and net-of-depreciation returns are more appropriate for such an interpretation. Equally, only 1 year's data are considered, as the primary focus of this note is the sensitivity of Thornton's results to calculation procedures and parameter values.

[8] Returns to capital in any year comprise profit in industry (the value of sales less turnover taxes and the industry wage bill) and depreciation charges. In this note, Thornton's calculations using net returns on capital are considered, although calculations using gross returns are also presented. The difference in the estimates obtained is small.

[9] This is estimated from employment, average wage, and social insurance deductions by industry.

[10] The capital stock data consists of fixed plus working capital. For 1960 official Soviet data are used, for 1964 nonofficial Soviet source data.

[11] The highest rate of return is for light industry (31.13 percent) and the lowest is for peat (3.62 percent). The average for all Soviet industry is 13.47 percent.

Recalculations of Thornton's Results Using Fixed-Point Methods

Thornton's estimates may be compared with recalculations obtained by similar methods used to reexamine some of the market economy efficiency loss calculations referred to earlier. A complete procedure must take account of the fact that demand conditions as well as supply conditions are needed to determine prices The method thus requires a specification of explicit demand functions in addition to the production functions All of the parameters of the model must be such that the state achieved by Soviet industry in the presence of differential charges can be replicated as an equilibrium state The differential charges are then removed and a new equilibrium computed In each case an equilibrium is characterized by an equilibrium set of prices such that at these prices demands equal supplies for all commodities, and no production process in any sector makes positive profits

Constant elasticity of substitution (CES) production functions are used for each sector of Soviet industry, the Cobb-Douglas case examined by Thornton is reexamined as a special case for which the elasticity of substitution in each sector is unity On the demand side, one consumer (the Soviet planner) has a utility function defined on the values added in each sector Thornton is concerned with a case where the mix of value added is preserved across Soviet industry, and this is replicated by using a CES utility function with a very low elasticity of substitution (0 001 is used) A further set of calculations is performed with Cobb-Douglas utility functions

The results of the recalculations are reported in table 1 For the case which most closely matches that considered by Thornton it can be seen that Thornton's estimates are not too far from those obtained using the fixed-point methods of calculation This closeness is not inconsistent with a finding by Shoven and Whalley (1972) that the approximation procedures for distortion calculations for market economies tend to move somewhat closer to the general equilibrium calculations as the elasticity of substitution rises, although larger discrepancies between methods of calculation have been found by Whalley (1973) in a model of the U.K economy For smaller values of the elasticities of substitution (say, 0 4 as reported in table 1) the discrepancies between estimates from various methods could well be more substantial If the mix of value added is constant, the index number problem of differences in the estimates of efficiency losses depending upon choice of price weights does not seem to be of major importance. This feature is perhaps surprising, as move-

TABLE 1

RECALCULATION OF THORNTON'S ESTIMATES OF STATIC EFFICIENCY LOSSES IN SOVIET INDUSTRY

ELASTICITY OF SUBSTITUTION IN PRODUCTION FUNCTIONS	CHANGE IN VALUE ADDED (%)	
	Valuation at Prechange Equilibrium Prices	Valuation at Postchange Equilibrium Prices
	Case A Constant Value-added Mix*	
0.4	1.33	1.32
0.5	1.64	1.64
0.75	2.42	2.41
1.0	3.13	3.12†
2.0	5.73	5.73
3.0	7.95	7.78
	Case B Cobb-Douglas Utility Function	
0.4	4.49	1.34
0.5	4.88	1.65
0.75	5.80	2.41
1.0	6.62	3.13
3.0	12.05	7.88

NOTE.—See text for detailed explanation.
* CES utility function with elasticity 0.001.
† Thornton's estimate (2.9) is actually of a saving in value of resources to produce the same outputs.

ments in factor and output (market clearing) prices as the distortions are removed are of substantial magnitude.

Along with the reasonable proximity of Thornton's estimate for particular parameterizations, the other striking feature of table 1 is the smallness of the estimates as the elasticity of substitution assumed for each of the sectors falls,[12] 0.4 corresponds to Weitzman's (1970) estimate of the elasticity of substitution in Soviet manufacturing using official Soviet data (he obtains a lower estimate using Western reconstructed data). With elasticities of substitution in this range the results in case A indicate that the static efficiency losses in Soviet industry might be 1.5 percent of value added rather than the 3 percent of value added that Thornton suggests.

The calculations reported in case B of table 1 allow for changes in the mix between values added (the Cobb-Douglas utility function calculations) and indicate that the index number problems from price changes can become severe in evaluating possible efficiency effects. Discrepancies of up to 300 percent are recorded in some cases The seriousness of this index number problem is reminiscent of the different growth rates observed for the Soviet economy by Moorsteen and Powell (1966) which depend upon the year for which prices are used [13]

---

[12] As the demand side of the economy is constrained in the calculations in case A such that the "consumer" purchases outputs in fixed proportions, it is perhaps not surprising that efficiency loss estimates fall in this case. The size of the difference between estimates is, however, substantial.

Although only recalculations have been presented here, it should also be pointed out that considerable elaboration on Thornton's calculations are possible within the computational framework used. On the production side of the economy it is possible, for instance, to incorporate an input-output matrix "on top" of the production functions This might be expected to "magnify" the curvature of the production possibility frontier and affect the efficiency loss estimates

References

Bergson, A "The Economics of Soviet Planning " New Haven, Conn Yale Univ Press, 1964

Gerschenkron, A , assisted by A Erlich "A Dollar Index of Soviet Machinery Output, 1927/28-1937 " RAND Corporation, R-197 Santa Monica, Calif RAND Corp , 1965

Harberger, A C "Efficiency Effects of Taxes on Income from Capital " In "Effects of Corporation Income Tax," edited by M Kryzaniak Detroit Wayne State Univ Press, 1966

Johnson, H G , and Mieskowski, P M "The Effects of Unionization on the Distribution of Income A General Equilibrium Approach " *Q J E* 84, no 4 (November 1970) 539–61

Moorsteen, R H , and Powell, R P *The Soviet Capital Stock 1928–1962* Homewood, Ill Irwin, 1966

carf, H E , with the collaboration of T Hansen *The Computation of Economic Equilibria* New Haven, Conn Yale Univ Press, 1973

hoven, J B , and Whalley, J "A General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U S " *J Public Econ* 1, no 3–4 (November 1972) 281–321

—— "General Equilibrium with Taxes A Computational Procedure and an Existence Proof " *Rev Econ Studies* 40, no 4 (October 1973) 475–89

ornton, J "Differential Capital Charges and Resource Allocation in Soviet Industry " *J P E* 79, no 3 (May/June 1971) 545–61

itzman, M L "Soviet Post-War Economic Growth and Capital-Labour Substitution " *A E R* 60, no 4 (September 1970) 676–92

alley, J "A Numerical Assessment of the April 1973 Tax Changes in the United Kingdom " Ph D thesis, Yale Univ., 1973

es, P J D *The Political Economy of Communism*. Oxford. Blackwell, 1962

This phenomenon has been termed the "Gerschenkron effect" due to an original ervation by Gerschenkron (1965) of the considerable sensitivity of growth-rate ures for output of Soviet machinery between 1928 and 1937 to the choice of year's e weights

Sharing Risks of Deferred Payment

Steven Shavell

Harvard University

Uncertainty often enters into an economic transaction when payment is deferred In this case the risk to transactors can usually be reduced by agreement on a plan that makes the amount paid depend on the occurrence of uncertain events

For example, if movements in the price level cannot be predicted, then the real value of a future payment is uncertain In order to eliminate this risk, some economists have suggested that payment should be made according to a cost-of-living escalator That is, payment should be a variable $(1 + \pi)N$ where π is the inflation in prices over the period in question and N is some nominal amount In principle, however, payment could depend in an arbitrary way on inflation The question then arises of what is a Pareto optimal payment plan—a plan designed so that no substitute arrangement could mutually benefit the transactors Pareto-optimal payment plans are not, in general, cost-of-living escalators They are determined jointly by transactors' attitudes toward risk, by opinions about the future rate of inflation, and by the correlation with inflation of income or outgo which is unrelated to the payment plan

Another example in which deferred payment involves risk concerns contracts made between parties who transact in different currencies under conditions of uncertainty over the future rate of exchange The usual type of contract, which fixes payment in one of the two currencies, is not Pareto optimal even if the parties engage in transactions in the forward exchange market A Pareto-optimal contract, on the other hand, links payment to the exchange rate in a way depending on a set of factors similar to those mentioned above for the case of inflation

Before discussing the two examples, Pareto-optimal payment plans will be briefly considered in a slightly more general context. The section below on these plans is most closely related to Borch (1960, 1962) on einsurance treaties and Arrow (1971) on optimal insurance policies but

---

This is a revised version of my paper "Optimal Payment Plans and Cost of Living Escalators" (M I T , June 1974) I wish to thank G Feiger, P Samuelson, J Yellen, and a referee for comments Financial support from the National Science Foundation and the Office of Naval Research is gratefully acknowledged

Journal of Political Economy, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

differs in its interpretation and in its emphasis on the effect of unrelated or "exogenous" changes in wealth on risk sharing.

Pareto-optimal Payment Plans

Suppose that transactor U, who is to receive payment, and transactor V, who is to make payment, each acts as if he were maximizing the expected value of a function-giving utility of wealth.[1] Let the utility functions of U and V be $U(\ )$ and $V(\ )$, respectively. Denote by $d_u(\cdot)$ and $d_v(\ )$ the densities of their subjective probability distributions over the real numbers, to be interpreted as states of nature.[2] Assume also that $d_u(s) = 0$ if and only if $d_v(s) = 0$.

A *payment plan* is therefore a function $x(\ )$ giving the amount $x(s)$ which V is to pay U when s occurs.[3] Suppose that $u(s)$ is the expected wealth of U exclusive of payment from V, given the occurrence of s, and that $v(s)$ is similarly defined. Then under a payment plan the expected utility of U is

$$\int U[u(s) + x(s)]d_u(s)ds \tag{1}$$

and that of V is

$$\int V[v(s) - x(s)]d_v(s)ds \tag{2}$$

(All these functions are assumed differentiable.) Accordingly, a Pareto-optimal payment plan $x(\ )$ maximizes (1) with (2) held constant.[4]

If neither transactor prefers risk, then a necessary and sufficient condition for Pareto optimality is that

$$d_v(s)V'[v(s) - x(s)] = kd_u(s)U'[u(s) + x(s)], \tag{3}$$

where k is a positive constant. This equation implies that the marginal rates of substitution between dollars in different states are the same for

---

[1] Of course, other arguments of the utility function (including what U transferred to V in the exchange preceding payment) are held constant and are suppressed in the notation.

[2] Consideration of the case in which the states of nature are vectors is straightforward but does not seem to add insight.

[3] Although as stated U and V have symmetric positions in the problem, we think of U as recipient, especially since in most real cases payment will be positive. If, as may sometimes be appropriate, it is assumed that payment is nonnegative, the consequent modifications (e.g., $\geq$ replaces $=$ in [3]) are, again, straightforward.

[4] If, say, U believes some event is possible and V does not, then there are no Pareto-optimal plans. This is because V would then be willing to contract to pay U any amount if the event occurred. The problem obviously does not arise if either the condition $d_u(s) = 0$ if and only if $d_v(s) = 0$ or some constraint on borrowing is assumed. We chose to make the former assumption in the interest of avoiding tedious consideration of cases in which constraints are or are not binding.

U and V The equation follows easily from an argument used by Arrow (1971, p 217) on optimal insurance policies or from the calculus of variations [5]

Now differentiate (3) to obtain an equation in $x'(s)$,

$$\begin{aligned} d_v'(s)V'[v(s) - x(s)] + d_v(s)[v'(s) - x'(s)]V''[v(s) - x(s)] \\ = k\{d_u'(s)U'[u(s) + x(s)] + d_u(s)[u'(s) + x'(s)]U''[u(s) + x(s)]\}, \end{aligned}$$

or, substituting for k and solving for $x'(s)$,

$$x'(s) = \frac{\begin{aligned} v'(s)[-V''(s)/V'(s)] - u'(s)[-U''(s)/U'(s)] \\ + \{[d_u'(s)/d_u(s)] - [d_v'(s)/d_v(s)]\} \end{aligned}}{[-V''(s)/V'(s)] + [-U''(s)/U'(s)]}, \quad (4)$$

where $U'(s) = U'[u(s) + x(s)]$ and similarly for $U''(s)$, $V'(s)$, and $V''(s)$

The slope of the payment schedule is therefore determined by the derivatives of wealth exclusive of payment, the levels of absolute risk aversion, and the difference in subjective beliefs as expressed by the proportional rate of change of the probability density Other considerations, such as bargaining power and market forces, place limits on the height of the payment schedule

The effect of the difference in beliefs is easy to interpret It is to insure that transactor in whose opinion the likelihood of higher s is growing most rapidly

In order to interpret other terms of the equation, assume that risk-averse U and V have identical opinions Then the equation implies that f $v'(s)$ and $u'(s)$ are of opposite sign, $x'(s)$ has the sign of $v'(s)$, offsetting movements in prepayment wealth for both transactors [6] But if $v'(s)$ and $u'(s)$ are of the same sign, $x'(s)$ is opposite in sign to the derivative of prepayment wealth for the transactor with the greater weighted level of isk aversion, offsetting movements in his prepayment wealth

Finally, note two simple facts which follow from the equation (or irectly from Jensen's inequality) (i) A payment plan for two risk-averse ransactors with identical beliefs and constant expected wealth exclusive f payment is Pareto optimal if and only if payment is constant. (ii) A ayment plan for a risk-neutral transactor and a risk averter with identical eliefs is Pareto optimal if and only if it leaves the risk averter with fixed nal wealth [7] This second case is approached as the ratio of degrees of sk aversion of the transactors increases Divide the numerator and differs in its interpretation and in its emphasis on the effect of unrelated or "exogenous" changes in wealth on risk sharing

---

[5] Borch (1960) first proved that an equation similar to (3) characterizes Pareto-optimal k sharing

[6] This does not mean that the final wealth of both transactors is necessarily smoothed n the sense, say, of a reduction in variance) although one might expect it to be the pical case

[7] Baily (1974) notes essentially this fact

Pareto-optimal Payment Plans

Suppose that transactor U, who is to receive payment, and transactor V, who is to make payment, each acts as if he were maximizing the expected value of a function-giving utility of wealth [1] Let the utility functions of U and V be $U(\ )$ and $V(\ )$, respectively Denote by $d_u(\ )$ and $d_v(\ )$ the densities of their subjective probability distributions over the real numbers, to be interpreted as states of nature [2] Assume also that $d_u(s) = 0$ if and only if $d_v(s) = 0$

A *payment plan* is therefore a function $x(\ )$ giving the amount $x(s)$ which V is to pay U when s occurs [3] Suppose that $u(s)$ is the expected wealth of U exclusive of payment from V, given the occurrence of s, and that $v(s)$ is similarly defined Then under a payment plan the expected utility of U is

$$\int U[u(s) + x(s)]d_u(s)ds \tag{1}$$

and that of V is

$$\int V[v(s) - x(s)]d_v(s)ds \tag{2}$$

(All these functions are assumed differentiable) Accordingly, a Pareto-optimal payment plan $x(\ )$ maximizes (1) with (2) held constant [4]

If neither transactor prefers risk, then a necessary and sufficient condition for Pareto optimality is that

$$d_v(s)V'[v(s) - x(s)] = kd_u(s)U'[u(s) + x(s)], \tag{3}$$

where k is a positive constant This equation implies that the marginal rates of substitution between dollars in different states are the same for

---

[1] Of course, other arguments of the utility function (including what U transferred to V in the exchange preceding payment) are held constant and are suppressed in the notation

[2] Consideration of the case in which the states of nature are vectors is straightforward but does not seem to add insight

[3] Although as stated U and V have symmetric positions in the problem, we think of U as recipient, especially since in most real cases payment will be positive If, as may sometimes be appropriate, it is assumed that payment is nonnegative, the consequent modifications (e g , $\geq$ replaces $=$ in [3]) are, again, straightforward

[4] If, say, U believes some event is possible and V does not, then there are no Pareto-optimal plans This is because V would then be willing to contract to pay U any amount if the event occurred The problem obviously does not arise if either the condition $d_u(s) = 0$ if and only if $d_v(s) = 0$ or some constraint on borrowing is assumed We chose to make the former assumption in the interest of avoiding tedious consideration of cases in which constraints are or are not binding

U and V The equation follows easily from an argument used by Arrow (1971, p 217) on optimal insurance policies or from the calculus of variations [5]

Now differentiate (3) to obtain an equation in $x'(s)$,

$$d_v'(s)V'[v(s) - x(s)] + d_v(s)[v'(s) - x'(s)]V''[v(s) - x(s)]$$
$$= k\{d_u'(s)U'[u(s) + x(s)] + d_u(s)[u'(s) + x'(s)]U''[u(s) + x(s)]\},$$

or, substituting for k and solving for $x'(s)$,

$$x'(s) = \frac{v'(s)[-V''(s)/V'(s)] - u'(s)[-U''(s)/U'(s)] + \{[d_u'(s)/d_u(s)] - [d_v'(s)/d_v(s)]\}}{[-V''(s)/V'(s)] + [-U''(s)/U'(s)]}, \quad (4)$$

where $U'(s) = U'[u(s) + x(s)]$ and similarly for $U''(s)$, $V'(s)$, and $V''(s)$

The slope of the payment schedule is therefore determined by the derivatives of wealth exclusive of payment, the levels of absolute risk aversion, and the difference in subjective beliefs as expressed by the proportional rate of change of the probability density Other considerations, such as bargaining power and market forces, place limits on the height of the payment schedule

The effect of the difference in beliefs is easy to interpret It is to insure that transactor in whose opinion the likelihood of higher s is growing most rapidly

In order to interpret other terms of the equation, assume that risk-averse U and V have identical opinions Then the equation implies that if $v'(s)$ and $u'(s)$ are of opposite sign, $x'(s)$ has the sign of $v'(s)$, offsetting movements in prepayment wealth for both transactors [6] But if $v'(s)$ and $u'(s)$ are of the same sign, $x'(s)$ is opposite in sign to the derivative of prepayment wealth for the transactor with the greater weighted level of risk aversion, offsetting movements in his prepayment wealth

Finally, note two simple facts which follow from the equation (or directly from Jensen's inequality) (i) A payment plan for two risk-averse transactors with identical beliefs and constant expected wealth exclusive of payment is Pareto optimal if and only if payment is constant (ii) A payment plan for a risk-neutral transactor and a risk averter with identical beliefs is Pareto optimal if and only if it leaves the risk averter with fixed final wealth [7] This second case is approached as the ratio of degrees of risk aversion of the transactors increases Divide the numerator and denominator of (4) by $-V''(s)/V'(s)$ and let the ratio of degrees of risk aversion approach infinity Then $x'(s)$ approaches $-u'(s)$ Similarly, if the ratio approaches zero, $x'(s)$ approaches $v'(s)$

---

[5] Borch (1960) first proved that an equation similar to (3) characterizes Pareto-optimal risk sharing

[6] This does not mean that the final wealth of both transactors is necessarily smoothed (in the sense, say, of a reduction in variance) although one might expect it to be the typical case

[7] Baily (1974) notes essentially this fact.

Two Examples

A Indexing and Uncertainty over the Rate of Inflation

If s is the rate of inflation over a period and wealth is interpreted as real wealth, then a payment plan gives for any s the real payment $x(s)$, the nominal payment is $(1 + s)x(s)$ Consequently, a payment plan $x(s) \equiv c$, a constant, represents a cost-of-living escalator, a plan with $x'(s) > 0$ implies that real payments actually increase with inflation, a plan with $x'(s) < 0$, such as a nominally fixed payment, implies that real payments do not keep up with inflation

If it were true that everyone shared the same beliefs about inflation and had no source of real income affected by it—that is, $u(\ )$ and $v(\ )$ constant—then a cost-of-living escalator would be Pareto-optimal (items i, ii) Hence, a situation in which all real income and wealth are protected against inflation is self-supporting in the sense that any new Pareto-optimal payment plan would be a cost-of-living escalator

For most transactors, however, opinions over the future course of inflation differ and real income exclusive of payment is correlated with inflation Therefore the results of the previous section show that a Pareto-optimal payment plan is rarely a cost-of-living escalator From the discussion following (4), it becomes intuitively obvious that under a Pareto-optimal plan payments are made as if to compensate for other inflation losses or gains For instance, if U's weighted degree of risk aversion is sufficiently great compared with V's, then real payment to U will rise with inflation if his real income from other sources is negatively correlated with inflation and fall if his income is positively correlated

[The random variables $u(\ )$ and $v(\ )$ describing income from other sources may depend on $x(\ )$ For instance, the portfolio of an individual would probably change if his wage income were protected against inflation Hours worked might also change We do not consider explicitly such effects here[8], we are concerned with risk sharing and therefore with what the *final* relationship must be among $u(\ )$, $v(\ )$, and $x(\ )$ for Pareto optimality]

We now comment briefly on Pareto-optimal linkage for two important payment plans wage contracts and bonds It seems reasonable to argue that for many individuals real disposable income would have fallen recently even if their wages had been fully escalated This is because average tax rates increase with nominal income and because, on balance, real income from asset portfolios—including cash, savings, and equity in insurance policies, pension funds, a home, or stocks—has often been negatively correlated with inflation If this is expected in the future and if firms may be assumed risk neutral (or at least much less risk averse than wage earners), then according to a Pareto-optimal escalator clause real wages should rise with inflation Of course, a firm would not agree to such a clause without getting something in return, wage earners ought to be willing to accept a lower expected real salary in order to enjoy extra inflation protection

---

[8] In unpublished notes, Blinder (1975) considers some of these effects in a context similar to that of this section

If such escalators are Pareto optimal, competition should enforce them The way this would work is that, for example, a firm which offered a fixed nominal wage or even a cost-of-living escalator to an individual whose other income was negatively correlated with inflation would have to pay a higher expected wage than a firm which offered an optimal escalator Similar remarks obviously apply when the Pareto-optimal escalator does not involve a greater than cost-of-living increase

Because wage earners differ, competition should encourage firms to offer a choice of escalators Some escalators would have a relatively low expected real wage and a high degree of escalation, others a relatively high expected real wage and little escalation A wage earner will wish for greater escalation the more averse he is to risk, the greater the amount of his nonwage income and the more negative its correlation with inflation, and the larger the degree to which his expectations of high rates of inflation exceed those of the firm (eq [4]) The choice of escalators permitted by allowing wage earners to mix according to any desired ratio a nominally fixed and some highly escalated contract would likely be sufficient as a practical matter

The case with bonds is similar, except that the predominant demand might not be for instruments as highly escalated as in the case of wage contracts This claim is based on casual empiricism—that the correlation with inflation of other income and wealth is not as unfavorable for the average bondholder as for the average wage earner In any event, to accommodate variable demand it would probably be enough for issuers to float only two bonds, one with a nominally fixed yield and one with linkage, for investors could purchase these in any proportion

Although the theoretical arguments of the present section would suggest otherwise, widespread linkage of payments to the price level has not been observed in this country One explanation, stressed by Blinder (1975), is that individuals are not sufficiently more risk averse than firms (This explanation relies on the existence of transaction costs [bargaining, legal, etc], for without these any difference in attitude toward risk would make some kind of linkage mutually advantageous) While there is obvious truth in this explanation and some evidence in its favor—for example, the "capping" of escalators in certain wage contracts—other not necessarily competing explanations also bear consideration. One of these has no doubt been the episodic nature and mild pace of inflation. Regarding wages, another explanation may be that most agreements in the labor market do not bind employees, presumably, employees could look for work elsewhere if payments under a plan turned out to be low [9] And, in the case of bonds, an explanation may be the unfavorable treatment under the tax laws of interest payment on indexed corporate debt.

Nevertheless, with the increased importance and awareness of inflation, a tendency toward linkage is to be expected Whether this tendency is socially desirable is not clear Our argument for Pareto optimality was distinctly microeconomic, concerned with only two transactors It took no account of the influence which extensive use of linkage would have on the economy as a whole through effects on individual consumption, portfolio behavior, and expectations When viewed in this larger context, linkage may not be in everyone's advantage [10]

B Payment in Foreign Currency and Uncertainty over the Exchange Rate

Consider a contract $x(\ )$ for deferred payment between risk-averse or risk-neutral parties U and V who transact in different currencies For simplicity, assume that both agree over the probability distribution F describing the future rate of exchange Let s be the number of units of U's currency purchasable with one unit of V's currency at the time of payment Then, if $x(s)$ is the payment in V's currency, $sx(s)$ is the payment in U's currency Consequently, the expected utility of U is

$$\int U[u(s) + sx(s)]dF(s) \tag{5}$$

and that of V is

$$\int V[v(s) - x(s)]dF(s) \tag{6}$$

Because U does not receive precisely what V gives up, the condition for a Pareto-optimal payment plan is a little different from that of the previous section An argument analagous to Arrow's shows that the condition for Pareto optimality becomes

$$V[v(s) - x(s)]/sU[u(s) + sx(s)] = k, \tag{7}$$

---

[9] One would therefore expect the following two factors to militate against indexed wage agreements (*a*) low cost of changing jobs, (*b*) inability to negotiate similar contracts for a large fraction of employees in an industry

[10] Fischer (1975) has analyzed the effect on portfolio behavior of the opportunity to buy index-linked bonds and has constructed an example in which the consequence of their introduction is to make everyone worse off

where k is a positive constant The meaning of the equation is that the terms of trade in utility remain constant over exchange rates Differentiation of (7) and solution for $x'(s)$ would yield an equation similar to (4) and with very much the same interpretation It is clear from (7) that a contract specifying a fixed payment in the currency of either transactor is not in general Pareto optimal Other things equal, a risk-averse U should receive less in his own currency when the exchange rate is in his favor [By other things equal is meant that $u(\ )$ and $v(\ )$ are constant] If $x'(s) > 0$, U's receipts of $sx(s)$ clearly increase with s,[11] if $x'(s) \leq 0$, equality in (7) cannot be maintained unless $sx(s)$ rises with s This is easy to understand, for it says that U receives less in his currency when it is costly for V in exchange for more when it is cheap

It is assumed above that the transactors themselves share the risk of exchange rate fluctuation When the forward rate or the implicit forward rate (associated with the spot exchange rate and the interest rates in the transactors' countries) make covering exchange rate risk very expensive, such an assumption may provide a good approximation to the truth[12] Under these circumstances, one guesses that many potential contracts are never consummated just because their terms (payment fixed in one currency) would not allow adequate risk sharing

When transactors use forward exchange markets, the common practice is for a contract to specify a fixed payment in, say, V's currency and for U to hedge against exchange rate fluctuation to the extent he desires by selling forward in V's currency It is easy to see that in this case also (7) will rarely be satisfied A complete analysis of Pareto-optimal contracts in the presence of forward exchange markets requires maximization of (5) over functions $x(\ )$ with (6) constant, where U and V each choose optimal forward transactions given $x(\ )$ The analysis, which will not be presented here, does not turn out to be interesting, the results amount to a complicated mathematical statement of the fact that Pareto-optimal contracts would make payment depend on the exchange rate not only to exploit the low "cost" of utility for U when s is high but also to offset variation of other income with changes in s Forward markets alone are not a sufficiently discriminating device for such shifting of risk

Concluding Comment

In the two examples above, payment plans were beneficial because insurance against price changes could not be purchased More generally, of course, opportunity for improving the allocation of risk when payment is deferred often arises when markets for claims contingent on states of the world are not complete However, it must be admitted that the scope for gains in welfare through use of payment plans is limited (though to a different degree) by the same two factors which frequently prevent formation of markets for contingent claims, namely, by transaction costs and moral hazard

---

[11] Here it is assumed that payment is positive

[12] The effective cost of forward transactions is likely to be higher the longer the duration of the contract and the larger the magnitude of the transaction

References

Arrow, Kenneth J "Uncertainty and the Welfare Economics of Medical Care " In *Essays in the Theory of Risk-Bearing* Chicago Markham, 1971

Baily, M N "Wages and Employment under Uncertain Demand " *Rev Econ Studies* 61 (1974) 37-50

Blinder, A S "Notes on the Choice of an Escalator Clause " Unpublished notes Princeton Univ , 1975

Borch, Karl "The Safety Loading of Reinsurance Premiums " *Skandinavisk aktuarietidskrift* (1960), pp 163-84

——— "Equilibrium in a Reinsurance Market " *Econometrica* 30 (1962) 424-44

Fischer, Stanley "The Demand for Index Bonds " *J P E* 83, no 3 (June 1975) 509-34

An Inquiry into the Economics of Race-Track Gambling

Arthur Gruen

Tufts University

"All horseplayers die broke " [DAMON RUNYON]

In the face of mounting fiscal pressures state governments are looking increasingly to gambling as a source of revenue If the sale of gambling is to provide revenues, however, it is important to understand the demand for this good in order to adopt the appropriate pricing mechanism Standard economic theory posits a model where the quantity of a good consumed is negatively related to the price of the good if we fix tastes, income, and the prices of other goods Such a concept is intuitively reasonable as well as useful in analyzing much of observed economic behavior Does gambling have a price in this sense, and do consumers behave in the same rational framework so that this good fits the model?

In this paper I will examine the demand for horse-race wagering Part I will present a model which attempts to systematically explain and predict the behavior of consumers who bet on horses In Part II a demand equation for wagering will be estimated From this the optimum price or rate of taxation can be derived so as to maximize the revenue received by both the state and the track

I

The most obvious feature of gambling is the involvement of risk In a andmark article on individual's reaction to risk, Friedman and Savage (1948, pp 279–304) rationalized ostensibly inconsistent behavior (the ame individual purchasing an insurance policy and a lottery ticket) by an application of orthodox utility analysis By introducing decision theory (Raiffa 1968) we can extend this analysis to incorporate the direct participation of the bettor in horse-race wagering

Let us approach this problem by considering a race with five horses

---

* The author is grateful to Professor Thomas F Cooley for his critical comments and encouragement with this project

Journal of Political Economy, 1976, vol 84, no 1]
© 1976 by The University of Chicago All rights reserved

TABLE 1

Five-Horse Race with Win Bets

Horse	Amount Bet ($)
A	50,000
B	10,000
C	20,000
D	30,000
E	40,000
Total	150,000

and only win betting allowed. Assume bets are made as shown in table 1 The total amount bet is known as the "handle " The state and the track each receive a percentage commission, or "takeout," on the handle which is set by law The remainder of the handle is returned to the bettors who picked the winning horse If horse A should win and we assume a total commission of 10 percent, the winning bettors receive $135,000, which represents a profit of $85,000 or a return of $1 70 per dollar Since bets are bought in $2 00 units the actual winnings would be $3 40 per bet (The posted payoff would be $5 40, which includes the $2 00 which is returned to the bettor) Each bet will affect the payoff, as payoffs are determined by the amount bet on each horse Yet no single individual would expect his or her own bet to change the odds, and thus each bettor is a price taker in a perfectly competitive market Actually, it is sufficient for the bettor to be confident that each payoff will fall within a predictable range in order to make a decision The parimutual boards provide this information continuously to the bettor until the race begins

The second bit of information required is the probability of each horse winning The posted odds only reflect the collective wisdom of all individuals who bet on the race It has no direct relationship to any "internal" probability In fact, no such internal mathematical probability exists and the wagerers must determine for themselves the "true" odds

The potential bettor can refuse to bet or bet on one or more of the five horses in the race One can obtain the expected monetary value (EMV) of each option by multiplying each possible outcome by its probability and summing these products over all possible outcomes The higher the expected monetary value of a given option the more bets the individual would be willing to make This immediately follows from marginal utility analysis A relative increase in the expected marginal utility per dollar of a given commodity will cause a redirection of expenditures to this commodity In the limit, if an individual is 100 percent sure that a horse will win, all possible resources should be used to make bets The crucial point is that bets are an increasing function of EMV

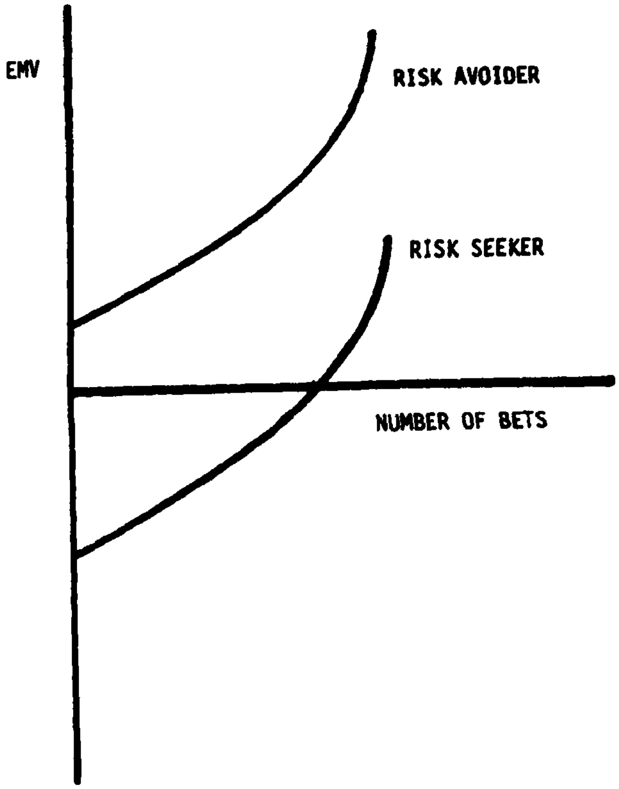


FIG 1

No assumptions about the nature of an individual's utility function or risk preference are required Two cases are illustrated in figure 1 [1] The slopes are increasing due to the resource constraint of the individual The risk seeker with an increasing marginal utility of income may be willing to bet even with a negative EMV It is also perfectly consistent for the risk avoider to make a bet given a high enough EMV

Let us assume that bettor 1 has analyzed the race and believes the probability of each horse winning is as shown in table 2 Given the probable payoffs with a 10 percent takeout, we can derive this bettor's EMV schedule Let us assume for simplicity that both bettors have a linear utility function so that the relationship in figure 2 holds This bettor would choose option 5 which has the greatest value, and bet on horse D. Option 4 also has a positive EMV, and thus a bet on horse C would be made as well

---

[1] This is implied by Rothschild and Stiglitz (1971, pp 66–85).

TABLE 2

EMV SCHEDULE OF BETTOR 1

Option	Bet	Probability of Winning (%)	Payoff ($)	EMV
1	No bet	0	0 00	00
2	A	30	3 40	– 38
3	B	5	25 00	– 65
4	C	15	11 40	01
5	D	25	7 00	25
6	E	25	4 60	– 35

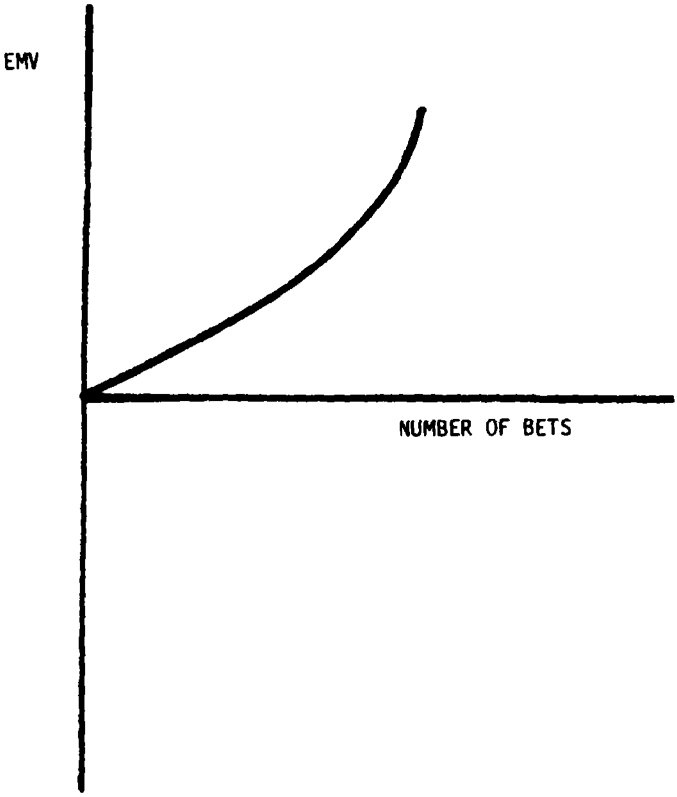


FIG 2

Bettor 2, however, has a different perception of the race and a different EMV schedule (see table 3) The highest expected monetary value is option 2, which indicates a bet on horse A. The next most feasible choice is option 1, which has a value of zero and represents no further betting

Let us now assume that the takeout has been increased to 17 percent

TABLE 3

EMV SCHEDULE OF BETTOR 2

Option	Bet	Probability of Winning (%)	Payoff ($)	EMV
1	No bet	0	0.00	.00
2	A	40	3.40	.16
3	B	5	25.00	-.65
4	C	10	11.40	-.66
5	D	20	7.00	-.20
6	E	25	4.60	-.35

TABLE 4

NEW EMV SCHEDULES

Bet	Payoff with 17% Takeout ($)	EMV for Bettor 1	EMV for Bettor 2
No bet	0.00	.00	.00
A	2.80	-.56	-.08
B	22.80	-.76	-.76
C	10.40	-.14	-.76
D	6.40	.10	-.32
E	4.20	-.45	-.45

The track now shares $25,500 with the state, and only $124,500 is returned to the bettors. This reduces the return on horse A to $1.49 per dollar which is rounded down to $1.40, and winnings become $2.80 on each bet instead of $3.40. The 18¢ which the bettor loses due to this rounding down process is known as "breakage." Nothing has happened to the horses themselves or to the relative payoffs, and thus there is no reason for the bettors to alter their perceptions. The only change is the amount of the takeout. Since all payoffs are now 7 percent less, the relative odds remain the same.[2] Table 4 summarizes each bettor's new EMV schedules, given the same race but with a 17 percent takeout.

Thus an increase in the takeout, holding everything else constant, results in bettor 1 betting only on horse D while bettor 2 does not bet at all. The total number of bets, therefore, declines with an increase in the takeout. What is the nature of this takeout? If there were none, the total handle would be returned to the bettors, and our problem would reduce to mere redistribution of wealth. Its existence, however, ensures that bettors in the aggregate leave the track with less than when they came in. They thus pay a premium in order to participate. We have established theoretically that an increase in this premium while holding all other factors fixed will result in a decrease in the number of bets. Thus this premium serves the economic function of the purchase price of a bet which, like other commodities, has a downward-sloping demand curve In the next section we shall test empirically this hypothesis and attempt to draw implications for optimum price policy.

---

[2] The rounding-down procedure may create minor distortions so that all payoffs might not be reduced by *exactly* the same percentage. The maximum amount of distortion, however, cannot exceed 20¢ per bet and is negligible.

II

The data to be used are the annual figures for Aqueduct Race Track and Belmont Park in New York City (*New York Racing Association Annual Report, 1940–1969*) The parimutual system was instituted in 1940, the first year explicit statistics were recorded, and our time series extends to 1969 A single location was chosen in order to correct as much as possible for tastes or customs, which may vary geographically. Furthermore, since the population and the volume wagered is so large relative to other locales, no single race or event is likely to distort the aggregate data and bias the results New York, along with Florida and California, is a racing center, which means that the highest purses and the best horses are attracted to these tracks No northeastern track can match New York in terms of quality of the product so the betting should not be affected by nearby competition The racing year is divided into meets which are held alternately at Aqueduct and Belmont The handle in the winter and spring meets can and does differ, therefore the New York Racing Association alternates the assignment of the meets from year to year so as not to systematically penalize one track By using annual data we avoid problems of seasonal variation The population sample is the area readily accessible to the two tracks the five counties of New York City plus Nassau and Westchester Counties

Our basic model postulates that the number of bets per race per capita is a function of the price of a wager plus real income per capita of the potential patrons We do not include admission fees because these are more properly the nominal cost to the patron of enjoying the facilities and the spectacle of the track itself quite apart from the opportunity to gamble The determination of the takeout is a political decision which the tracks are not free to alter in response to demand We can also regard income per capita as an exogenous variable Since the horse-racing industry is such a miniscule fraction of the economy of the New York Metropolitan area, changes in the income of this industry would have no noticeable effect on the total income of the region Thus our explanatory variables are exogenous to the model and independent of each other

The peripheral costs of racetrack betting such as fees, taxes, required attendance, and lack of credit make it the most expensive mode of gambling For those individuals who already have chosen this alternative it is unlikely that small changes in these peripheral costs will affect the bettors We therefore expect the potential substitution effect to be relatively constant over time. Off-track betting (OTB), on the other hand, would have a significant effect on on-track gambling. It is for this reason that our sample period stops before OTB was instituted. Gimic wagering (e g , picking the first three horses in exact order) has the allure of huge payoffs and is designed to promote betting Again our sample period stops before such bets were instituted These factors can and should be incorporated into a forecasting model (Coate and Ross 1974), but this is beyond our present scope The daily double has always been available and thus is a constant factor over the sample period

In analysis of stock-market behavior mention inevitably is made of waves of optimism or pessimism which affect stock prices above and beyond the strictly economic factors such as profits or price-earnings ratios In gambling it is just as reasonable to assume that people will bet more if they are optimistic about the future and less if they are pessimistic Part of this will be picked up by an increase in real per capita income, but bettors may be in a more expansive mood than can be explained by income alone In a depression, of course, expectations of hardship may curtail spending by a greater amount than merely the effect of lowered income There are several ways this phenomenon can be measured One way is to consider national unemployment rates Economists are generally agreed that unemployment rates lower than 4 percent (at least during this sample period) represent a full-employment economy Use of a dummy variable during years of economic boom will pick up this effect of optimism while avoiding correlation with income At the same time, it is reasonable to use a dummy variable from 1940–1944 to account for the depressing effect on consumption of rationing and wartime sacrifice At no other time during the sample period did the national unemployment rate climb much over 7 percent

Our variables are as follows

BETS = Number of bets per race per capita or "quantity "[3]

P = Percentage of the handle which is not returned to the bettors This consists of the state commission and the share to the track plus the breakage In our analysis it is the "price" of the bet

Y = income per capita deflated by the Consumer Price Index in 1967 dollars

PES = 1, 1940–44, 0, otherwise, depression or pessimism factor

OP = 1 when the unemployment rate is less than 4 percent, 0, otherwise, optimism or prosperity factor

---

[3] In the context of this model, each bet is a purchased commodity with a given price d *not* a dollar amount expended for the purchase of some other commodity BETS are easured in real units and thus no deflation is needed

A linear demand curve in all the variables yielded the best fit[4]

$$\text{BETS} = 0226 - 1475\text{X}10^{-4}\,\text{P} + 4118\text{X}10^{-5}\,\text{Y} - 0087\,\text{PES} + 0019\,\text{OP}$$

t-statistic	(5 18)	(4 93)	(7 81)	(5 70)	(3 49)
elasticity		(1 57)	(98)		

R^2 (corrected) = 814 $F(4,25)$ = 32 66 D-W = 1 6

At the 2 5 percent significance level we can accept the hypothesis that there is no autocorrelation The income effect is positive, which indicates that bets are a normal good while the income elasticity of demand is higher than many might expect The optimism and pessimism variables have the correct sign, and though it may be possible to find cleverer proxies these seem to serve adequately enough The crux of our theory centers on the price variable which is found to be negative and significant The theory predicted just this result, and it supports the view that a bet is like any other consumption good We find further that not only do bettors react to price, but at the rates set by the state the demand is actually price elastic[1] Since there are so many substitute forms of entertainment even aside from gambling, on reflection this should not be too surprising a result If this is the case, therefore, a decrease in the takeout should result in an increase in revenue Since price is the percentage of each bet retained by the track and the state, maximizing the product of this percentage and the number of bets will maximize the total amount which both receive Since we have a downward-sloping demand curve with respect to price, the price at which the elasticity of demand is unity will yield the maximum revenue For example, let us consider 1969, the final year in our sample We fix the curve for 1969 by setting

$$\begin{aligned} \text{Y} &= \text{Y}(1969) = 4722\ 22 \\ \text{PES} &= \text{PES}(1969) = 0 \\ \text{OP} &= \text{OP}(1969) = 1 \end{aligned}$$

Thus BETS = 0439 − $1475\text{X}10^{-4}$ P which becomes P = 2976 − 67796 BETS The point of unitary elasticity is at the midpoint of the demand curve so the optimum price is 14 88 percent The actual takeout in 1969 was 17 16 percent By enacting no other policy changes other than merely lowering the takeout by 2.28 percent the state and the track could have increased their net proceeds by $11,136,000

---

[4] Since betting increases faster than income, Y^2 was tried for this equation The results were similar, but this form was rejected because the degree of significance of the income variable was substantially reduced and the fit was not as good Various other linear and nonlinear forms were also tried The equation reported, however, proved to have the best fit and was most useful in determining an optimum price

Race-track wagering illustrates the fallacy of composition While it is true that bettors in the aggregate are destined to lose, this does not imply that for any *single* bettor the act of gambling entails a probabilistic loss and is irrational In fact we have demonstrated that those individuals who decide to go to the track and bet do indeed behave in a rational, price-elastic, utility-maximizing fashion The failure of the state to realize this fact is costing it needed revenue Each year the state can determine the takeout percentage for the coming year in order to maximize its potential revenue All that is required are estimates of income Thus we have devised a model where through mere price manipulation—virtually a costless procedure—we have a method for enriching both the profits of track owners, which should guarantee their cooperation, and the treasuries of state governments

References

Coate, D , and Ross, G "The Effect of Off-Track Betting in New York City on Revenues to the City and State Governments " *Nat Tax J* 27 (March 1974) 63–69

Friedman, M , and Savage, L J "The Utility Analysis of Choices Involving Risk " *J P E* 56, no 4 (August 1948) 279–304

New York Racing Association Annual Report New York New York Racing Assoc , 1940–69

Raiffa, H *Decision Analysis Introductory Lectures on Choices under Uncertainty* Reading, Mass Addison-Wesley, 1968

Rothschild, M , and Stiglitz, J E "Increasing Risk II Its Economic Consequences " *J Econ Theory* 5 (March 1971) 66–85

Currency Denominations and the Price Level

Chau-nan Chen

Academia Sinica and National Taiwan University

I

The need for issuing bigger denomination notes arises every 5 or 10 years in an economy like Taiwan where nominal income has been growing tremendously rapidly and where the public's currency holdings loom large in its total stock of money[1] The authorities, however, have been reluctant to meet the needs of trade on the ground that the issuance of big-denomination notes is likely to be interpreted by the public as a sign of the overissue of currencies, and the mere expectation of inflation may drive the price level up even though the total stock of money is held onstant On the other hand, there was recently a shortage of 1-yuan oins in Taiwan The authorities, in this case, responded very promptly nd effectively in meeting the difficulty of coin shortages by issuing 1-yuan otes first and then by managing to provide an adequate amount of new ins[2]

Introduction of bigger denomination notes to relieve the public from e inconvenience of a higher price and unchanged maximum denomation implies a shift to a better denomination mix of money (currency) y the same token, mitigating the difficulty of coin shortages and proding an adequate amount of small-denomination currencies also imply improvement in the currency size mixture The purpose of this note to analyze the macroeconomic effects of shifts in currency size mixture

---

am much indebted to Robert Barro for his most inspiring and helpful comments, torial as well as substantive I would also like to thank Harry G Johnson, Rudiger rnbusch, and the referees for their stimulating suggestions Research support from the tional Science Council of the Republic of China is also gratefully acknowledged

Need for bigger notes, in an effective sense, seems to depend not so much on real wth as on price inflation With real growth, the need for bigger notes arises only if it olves a larger unit value of transactions Throughout this paper I shall assume that growth is "neutral"

The premium is not necessarily paid for the metallic content Fisher (1926) reported stomers of retail stores often found it impossible to make their purchases because they ed the small denominations necessary, and because the storekeeper lacked the same ll denominations, and could not make change To meet the difficulty, 30,000,000 of e notes were issued, and they were so much in demand that dealers paid a premium them" (p 114)

nal of Political Economy, 1976, vol 84, no 1]
76 by The University of Chicago All rights reserved

Specifically it will be shown that an improvement in the size mixture of money would be contractionary on the price level

II

An improvement in the size mixture of money would have three types of effects on the real demand for money First, the real demand for money would be reduced since fewer real units of the more efficient money are now required to handle a given job—that is, a given real flow of monetary expenditure This effect is the efficiency aspect of an improved size mixture of money Second, the demand for real cash balances would be increased because the more optimal size mixture implies that money has become more attractive Substitution of money for barter exchange, and of money holdings for other asset holdings, including commodity inventories, would occur These shifts, which involve an increase in the real balance to income ratio, can be called the "substitution effect" Finally, the real flow of income would be increased to the extent that the increased use and efficiency of money implies a saving of labor and other real resources On this account there would be a positive income effect on the demand for money

Under today's fractional reserve banking system there is an additional contractionary force at work A better denomination mix of currency will lead to a shift out of deposits into currency—that is, a rise in the currency to deposits ratio, which, in turn, will reduce the money multiplier With an unchanged nominal quantity of high-powered money, the nominal stock of money will decline, rendering an additional contractionary force on the price level This can be called the "tight money effect" Thus for an improvement in the currency denomination mix to be inflationary, the direct efficiency effect has to be so strong that it outweighs the combined forces of the substitution effect, the income effect, and the tight money effect

Historical evidence, however, seems to suggest that the efficiency effect is not even strong enough to offset the substitution effect Although the "efficiency" of money has almost surely increased over the past decades, the income velocity of money has declined secularly in most of the countries This seems to suggest that the substitution effect of an increased efficiency of money has outweighed the direct efficiency effect[3]

III

That the substitution effect is more likely than not to outweigh the efficiency effect can also be analyzed within the framework of general equilibrium (Tobin 1969, 1970)

---

[3] This is a supplementary explanation of Milton Friedman's well-known thesis that money is a luxury good

To abstract from the income and tight money effects, assume that real output is fixed and that the total stock of money consists solely of hand-to-hand currencies of different denominations Let S and S^* stand for the demand for and the supply of small-denomination currency Let B and B^* stand for the demand for and the supply of big-denomination currency Finally, let P denote the nominal price level in terms of big-denomination notes, initially unity, and r the premium (greater or less than one) on the small-denomination notes When the actual proportion of supplies between the two denominations deviates from the desired proportion, the exchange rate between them would tend to vary But technical and institutional constraints prevent the currency exchange rate from adjusting adequately [4] This inadequate or imperfect adjustment of the exchange rate is the premium Thus, r represents not only the exchange rate (the price of small notes in terms of big notes) but also the additional user cost of the small-denomination currency Accordingly, P/r is the nominal price level in terms of small-denomination currency and $G = aP + (1 - a)P/r$ with $a = B^*/(S^* + B^*)$ is the general price level A highly simplified portfolio-balance model can be expressed by the following two equations

$$S(r)\ P = rS^*, \tag{1}$$

$$B(r)\ P = B^* \tag{2}$$

The nominal cash balances in both equations are expressed in terms of big-denomination currency [5]

Starting with an optimal size mixture of currency denominations—that is, S^* and B^* set in proper proportion so that $r = 1$—consider what would happen if there occurs a shortage of small-denomination coins [6] Coin shortages with an unchanged quantity of money (currency) stock mean that S^* decreases while B^* increases by the same amount (dA), that is,

$$-dS^* = dB^* = dA \tag{3}$$

---

[4] Partly due to limited divisibility of currency denominations and partly due to atutory restraints

[5] The nominal cash balances in the two equations can also be expressed in terms of nall-denomination currency

$$S(r)\ P/r = S^*,$$
$$B(r)\ P/r = B^*/r,$$

at they are essentially no different from (1) and (2) in the text

[6] In a nonshortage situation, the total money supply, $S^* + B^*$, is set, but the oportion of supplies between the two denominations is allowed to vary to get $r = 1$ ll these values S° and B° Then $S^* < S^{\circ}$ (for the same $S^* + B^*$) implies $r > 1$ For ical consistency, the premium has to be defined as the additional user cost of the nomination in shortage

Differentiating (1) and (2) totally, using the relationship (3) and solving, we have

$$dr/dA = (S^* + B^*)/\Delta, \tag{4}$$

$$dP/dA = (S^* - S_r - B_r)/\Delta, \tag{5}$$

$$d(P/r)/dA = -(S_r + B_r + B^*)/\Delta, \tag{6}$$

$$dG/dA = -(S_r + B_r)/\Delta, \tag{7}$$

where $\Delta = B_r S^* - S_r B^* + S^* B^*$

If it is assumed that small currency, big currency, and all the other assets excluded from the two balance equations are gross substitutes, that is,

$$S_r < 0, \qquad B_r > 0, \tag{8}$$

it then follows from the theorem on gross substitutes that

$$-S_r > B_r \tag{9}$$

With this amount of information on the signs and relative magnitudes of S_r and B_r, we know immediately from (4) that

$$dr/dA > 0 \tag{10}$$

Coin shortages cause the coins to carry a premium

It is also apparent from (5) that

$$dP/dA > 0 \tag{11}$$

The increase in the average user cost per unit of money reduces the demand for cash balances while the "appreciation" of the small-denomination currency increases the quantity of money in terms of big-denomination currency Accordingly, there is an increase in the nominal price level in terms of big-denomination currency

The effect on the price level in terms of small-denomination currency is uncertain While the increase in the average user cost per unit of money does reduce the demand for cash balances, the "depreciation" of the big-denomination currency reduces the quantity of money in terms of small-denomination currency From (6) we know that

$$d(P/r)/dA \gtrless 0, \qquad \text{according as } -(S_r + B_r) \gtrless B^* \tag{12}$$

The general price level, however, would surely go up The quantity of money deflated by the general price level is not affected by the change in the currency exchange rate, but the demand for cash balances is reduced because of the increase in the average user cost per unit of money The excess supply of money has to be eliminated by an increase in the general price level

$$dG/dA > 0 \tag{13}$$

When a particular denomination is completely (rather than partially) deleted, the above analysis can remain valid simply by assigning r as a "shadow" premium on the denomination deleted We can therefore conclude that a shortage or an absence of a desirable denomination would be inflationary [7] Put another way, either issuance of bigger denomination notes to accord the needs of trade or provision of an adequate amount of small-denomination currency would be contractionary on the price level

The key to the above result is the assumption that small currency, big currency and all the other assets are gross substitutes The assumption is quite natural and reasonable, for, without this assumption one may obtain the perverse result that the denomination in shortage would carry a negative premium, as can be easily seen from (4) [8] It is now clear that the dominance of the substitution effect over the direct efficiency effect implies nothing more than the gross substitutes assumption

References

Chen, Chau-nan "Bimetallism Theory and Controversy in Perspective " *Hist Polit Econ* 4 (Spring 1972) 89–112

——— "Flexible Bimetallic Exchange Rates in China, 1650–1850 A Historical Example of Optimum Currency Areas " *J Money, Credit, and Banking* 6 (August 1975) 357–76

Fisher, Irving *The Purchasing Power of Money* Rev ed New York Macmillan, 1926

Tobin, James "A General Equilibrium Approach to Monetary Theory " *J. Money, Credit, and Banking* 1 (February 1969) 15–29

——— "Deposits Interest Ceilings as a Monetary Control " *J Money, Credit, and Banking* 2 (February 1970) 4–14

---

[7] Notice that

$$\begin{aligned} dr/dA &= (S^* + B^*)/S^*B^* \\ dP/dA &= 1/B^* \\ d(P/r)/dA &= -1/S^* \\ dG/dA &= 0 \end{aligned}$$

if $S_r = B_r = 0$, that is, if the currency exchange rate is perfectly flexible as in the case of flexible bimetallic exchange rates in China (Chen 1975) Flexible domestic exchange rates ensure the stability of the average purchasing power of money

[8] Elsewhere I (Chen 1972) have demonstrated that the gross substitutes assumption is a sufficient condition for Gresham's Law to be operative

Book Reviews

Economic Regulation of Domestic Air Transport Theory and Policy By George W Douglas and James C Miller III Studies in the Regulation of Economic Activity, no 10
Washington, D C Brookings Institution, 1974 Pp xii+211 $8 95

During 1970 and 1971 the authors of this book played active roles as expert witnesses in presenting the U S Department of Transportation's (DOT) case before the Civil Aeronautics Board (CAB) in the Domestic Passenger Fare Investigation (DPFI) This book presents the research findings and analyses supporting their recommendations to the DOT in this important proceeding Many of these recommendations were adopted by the DOT, and several of them were later accepted by the Board in its various decisions in the DPFI

Fundamental to their policy recommendations is the *authors'* belief that "economic efficiency is the overriding policy objective" (p 62) and that the public interest should be redefined to mean economic efficiency (p 182) In order to evaluate the effects of CAB regulation relative to economic efficiency they propose a model of industry behavior under economic regulation in which they "adopt the pragmatic view that CAB regulatory policy is or can be exogenous" (p 39) and that the industry moves toward a market equilibrium in response to the CAB's changing policies and constraints Since the CAB has full power over price and entry/exit, but only limited power over service quality, this model implies that the regulated airlines adjust to changes in CAB pricing policies by modifying the quality and quantity of their services

Recognizing that the CAB has failed to allocate traffic and revenues wherever two or more rival airlines serve a city pair, Douglas and Miller's model of the regulated airline industry states "that in any given market, and thus in the aggregate, whenever actual load factor rises above the break-even level and excess profits are earned, carriers will commence scheduling additional capacity This has the effect of driving the actual load factor down toward break-even" (and vice versa) "Thus, over time, actual market load factors will tend to gravitate toward break-even levels, where the break-even load is determined by two parameters (a) the structure and level of airline capacity costs and traffic costs, and (b) the fares approved and enforced by the Board Thus, in market equilibrium the level of per-passenger cost is determined by the fare chosen If the fare is lowered or raised, then per-passenger costs will fall or rise correspondingly" (p 40)

Douglas and Miller specify three aspects of economic efficiency in the airlines First, technical efficiency concerns "achieving any rate of output of a given quality at lowest total cost" (pp 72, 63). Second, allocative efficiency "refers to the optimal production (and consumption) rates of output of given quality, given the technically feasible lowest-cost relationship and given consumer demands for the outputs" (pp 73, 63) Third, service-quality efficiency, which requires the value of increased service quality to just equal the marginal cost of this increment in service quality so that an optimal price-quality combination obtains with price being cost based (pp 74, 68). They believe that traditional profit motives among the regulated airlines result in a reasonable degree of technical efficiency, and the airlines' ability to adjust service quality in response to regulated prices results in a fair approximation of allocative efficiency Thus, they assert that the major responsibility of the Board in promoting efficiency "is in the choice of the optimal price-quality combinations" (p 72)

Building on these theoretical and normative foundations, Douglas and Miller devote chapters 6 and 7 to demonstrating that the existing price-quality combinations resulting from CAB regulation have been inefficient due to the Board's failure to adopt the optimal price-quality level and structure, peak load pricing, and congestion pricing while imposing restrictions on charter operations and promoting cross-subsidization and discount (discriminatory) fares Furthermore, they find the CAB inefficient in essentially prohibiting the entry of new airlines into scheduled passenger service, limiting entry into individual city pairs, restraining exit, authorizing mergers that seldom result in significant cost savings, and allowing various collusive agreements among the airlines, such as capacity-limitation agreements Unfortunately, while Douglas and Miller are able to use their model to demonstrate that present price-quality combinations are inefficient, they fail to use it to provide examples of optimal combinations for even a few actual city pairs (pp 87–92)

The brief discussion in chapter 8 of the DOT's and the authors' roles in the DPFI emphasizes their success in having the Board recognize how it affects service quality when authorizing fare levels and structures From this the Board came to understand the importance of incorporating a passenger load factor standard as part of its fare decisions Also, the authors believe the CAB now correctly places more emphasis on costs of service in establishing fares and has moved to reduce the use of discount fares, a form of inefficient value-of-service rate making At the same time, however, the Board rejected the DOT's recommendation that airlines be allowed to set fares anywhere from 15 percent below to 15 percent above those calculated from a Board-approved fare formula

Finally, in their last chapter Douglas and Miller estimate that in 1969 the deadweight welfare loss from trunk carrier operations due to the nonoptimal price-quality level and structure associated with CAB-approved fares was between $248 million and $356 million, or 3 8–5 5 percent of the trunk carriers' $6,514 million in domestic revenues for that year (p 172) They also make a series of policy recommendations regarding actions the Board should or should not take to increase efficiency under existing legislation, they recommend regulatory reforms requiring legislative action, and they suggest deregulation be considered (pp 179–83)

Douglas and Miller adopt an extreme position when they define the public interest as economic efficiency This is compounded by their intentionally excluding public goods and externalities (such as national defense and pollution) from their analysis (pp 61–62) Even if Congress should accept efficiency as the overriding public policy objective, it seems doubtful that it would also adopt such a narrow definition of the concept

Given the authors' definition of economic efficiency, it is necessary to realize that they advocate achieving this efficiency through the use of a nondiscriminatory single price for each class of service in each city pair (p 65) In addition, they take the extreme position that price must equal long-run marginal *and* average costs for efficiency to be achieved in equilibrium rather than allowing price to be equal to long-run marginal cost but greater than long-run average cost (p. 63) In essence, they advocate airline performance comparable with that achieved in a purely competitive market structure, but with price set by a regulatory commission (or by the market if the optimal price diverges from the Board's fare formula by no more than plus or minus 15 percent).

Since the authors find no significant economies of scale in trunk carrier operations, they are able to conclude that the long-run marginal and average cost curves are horizontal and congruent so that both can be equal to a single price This finding, however, is based on output being measured in terms of available ton-miles or available seat-miles At the same time the authors recognize that average costs per revenue passenger-mile (another measure of output) decrease with increases in load factor (pp 13–14), and thus are not necessarily constant over relevant ranges of output Therefore, different slopes for average and marginal cost curves are obtained for different output measures, and these differences carry over into the efficiency criterion The authors seek to solve this problem by assuming an optimal service-quality load factor for fare-making purposes, thereby making average and marginal costs equal regardless of how output is measured Unfortunately, it is not clear that a single optimal load factor actually exists since other aspects of service quality can counteract the effects of different load factors For example, at some price passengers may be indifferent between flying on a new-type aircraft at a high load factor and on an older plane at a lower load factor If there is no single optimal load factor and if revenue passenger-miles provide the economically relevant measure of airline output, as appears to be the case,[1] long-run average costs need not equal long-run marginal costs for most output levels In such situations they cannot both be equal to a single price which, of course, means that in relevant situations the fundamental requirement of the Douglas/Miller efficiency criterion cannot be met All of which argues against the nondiscriminatory single-price requirement

Implicit in this book is the position that economic efficiency should be achieved in ways that benefit consumers The authors fail to mention that economic efficiency can also be increased from present levels by the CAB promoting a perfect cartel with the full pooling of revenues and profits, with some optimal quality of service being produced by the lowest-cost carriers and with passengers being offered a combination of second- and third-degree discriminatory prices designed to capture their consumer surplus. The CAB-regulated airlines have long been moving in this direction Their adoption of capacity agreements has improved productive efficiency within the cartel, and the CAB fare formula implemented in 1969 (with its fixed terminal charge and decreasing rates per mile for sequential mileage blocks) is an obvious example of second-degree price discrimination from which discounts are given through promotional fares (family plan, excursion, student, no frill, etc) One can understand why the authors might not desire this solution to the efficiency problem, but given *their* chosen criterion it warrants analysis [2] Actually, what the authors propose is economic efficiency *and* consumer benefit Many would agree with such a compound policy objective, but one could well expect the airlines, their employees and stockholders, the aircraft manufacturers and other suppliers to prefer economic efficiency and

---

[1] An analogy in the education industry to the airlines' using available seat-miles as a relevant measure of output would be for a professor to claim to be more productive because he lectured in a 300-seat hall rather than a 75-seat room, even though there were only 50 students present in each case Universities have long favored using student contact-hours (rather than seat contact-hours) as a rough measure of teaching productivity, but the airline industry has yet to come to a similar conclusion.

[2] The authors are aware of this alternative solution to the efficiency objective (see Bilas and Miller 1971)

producer benefit. The normative arguments stemming from this difference in opinion are obvious and are not resolved by being ignored.

Underlying the authors' contention that airline prices should be based on costs is the assumption that demand for airline services will continue its secular growth. There is some tentative evidence that changes are occurring in the airlines' growth rate, and one can conceive of secular stagnation and eventual decline in demand. It is not clear that cost-based fares (as defined by the authors) will be viable in such a situation, let alone efficient, yet this possibility is not considered in the book. Overall, the failure to include demand factors in price determination is a weakness in the analysis.

Another problem with Douglas and Miller's cost-based fares is that they propose their being derived from the costs of the existing regulated airlines, thereby implicitly including regulatory costs in these fares. Evidence from the relatively unregulated California intrastate carriers indicates the costs of regulation are substantial with output per employee of regulated airlines being about half that of unregulated carriers (see Jordan 1970, pp. 210–20). Furthermore, the closure of entry by CAB regulation probably allows labor unions to increase wage levels. Douglas and Miller recognize this latter possibility (p. 141) but fail to incorporate it into their analysis. Had they excluded the costs of regulation from their calculations they might not have concluded that current technical and allocative inefficiencies are small and that the inefficiency from nonoptimal price-quality combinations is as little as 3.8–5.5 percent.

Douglas and Miller have made an important contribution to a theory regarding the effects of CAB regulation by postulating an inverse relationship between fares and average load factor in a convincing manner. Based on this model, however, they assert that "(e)ssentially, then, the airlines should be relatively indifferent to price, at least within the feasible range" (p. 170). This may be so, but they fail to recognize that the regulated airlines obviously are not indifferent to price. Indeed, the industry's history has been one of repeated requests for fare *increases*. This implies that there is more to the problem than is encompassed by the model. For example, the regulated airlines may not consider prices to be determined exogenously by the CAB and thereby beyond their own control.[3]

Another contribution of this book is its demonstration that forcefully presented economic analysis can influence policy decisions within the executive branch of the federal government and in regulatory commissions. The intervention of various executive departments in regulatory proceedings is an important development, and this book provides a beginning to understanding what may be achieved by such activities.

William A. Jordan

York University, Toronto

References

Bilas, Richard A., and Miller, James C. III. "Pigou's Three Degrees of Price Discrimination." *Rivista Internazionale Sci. Econ. Commerciali* 18, no. 9 (1971): 1–11.

Jordan, William A. *Airline Regulation in America.* Baltimore: Johns Hopkins Press, 1970.

---

[3] On pp. 39 and 145 n. the authors argue that prices are set exogenously by the CAB, while on pp. 41, 131, 139–40, 153, 163, and 168 they indicate the carriers play an important role in fixing prices.

A Model of Macroeconomic Activity Vol. 1 · *The Theoretical Model*. By RAY C. FAIR. Cambridge, Mass. · Ballinger Publishing Co. 1974. Pp. xvi + 226. $13.50.

One may choose from among at least three possible approaches for constructing and describing a macroeconomic model The first is to build a very small model that gives the reader a good intuitive grasp of the essentials This approach lends itself to analytical analysis but has the disadvantage of giving a very rough picture of the way the world operates. A second approach is to build a very elaborate econometric model Such a model adds realism by incorporating a large number of the sectors of the real world and many of the complex interactions among these sectors On the other hand, an intuitive grasp of the inner workings of a model so complex is very difficult to acquire. One can, however, employ an intermediate approach by building a model complex enough to incorporate the most important sectors of the economy, yet small enough to be able to determine how changes in the endogenous variables (that occur as a result of a change in one or more of the exogenous variables) are based on the assumptions underlying the specification used

The author of this book uses the third approach His aim is to construct a model that is based on microeconomic foundations, that is sufficiently general to include the most important behavioral sectors of the economy, and that does not require the behavioral units to have perfect foresight Unfortunately, the resulting model is too complex to be analyzed by standard analytical methods. The author must, therefore, resort to writing a computer algorithm to simulate the model I must say that, after reading the book, I am again impressed with Ray Fair's ability to work with and simplify complex models, although, as will be indicated below, I do not feel that he exploited his model to the utmost

This book has eight chapters and an appendix The first, a general introduction, indicates the author's purpose in building the model, outlines it, and discusses the methodology This is an excellent chapter and well worth reading even if one does not intend to work through the equations and tables of results further on The author traces the relationship between his model and the work of others and then provides a very fine summary of the model, quoted below The model consists of five basic sectors banks, firms, households, a bond dealer, and the government In Fair's words

> At the beginning of each period each bank, firm, and household, knowing last period's values, receiving in some cases information from others regarding certain current-period values, and forming expectations of future values, solves an optimal control problem The model is recursive in the sense that information flows in one direction from the bond dealer, to banks, to firms, to households Banks, for example, are not given an opportunity to change their decisions for the current period once firms and households have made theirs After all decisions have been made at the beginning of the period, transactions take place throughout the rest of the period [P 4]

A discussion of methodology dominates the remainder of the chapter. Not only does Fair give an overview of how the model is solved, but he also discusses the sources of inaccuracy in the study. One of these is the conversion of the stochastic control problems into deterministic control problems through an operation that replaces the stochastic variables by their expected values The author notes that this procedure is completely accurate only when the model is linear, his model is not. Another source of inaccuracy is that the algorithm does not guarantee that the true optimum will be found These types of examples make the book a very valuable source for those wishing to learn some lessons about model building.

Two versions of the model are presented. In the first, the "noncondensed" version, Fair postulates two households, one a creditor and one a debtor, and two identical banks and firms He specifies more than one firm and one bank in order to demonstrate how an individual bank's or firm's behavior is affected by its assumptions about the future behavior of its competitors This version is solved by determining the solution of the optimal control problem for each behavioral unit and then computing the transactions that take place between them. The other, the "condensed" version, aggregates the banks and firms to yield a bank sector and a firm sector The behavior of the banks, households, and firms in this version is approximated by simple equations and algorithms In chapters 2–5, which treat the behavior of the various individual sectors, the "noncondensed" model is presented first, with the "condensed" version briefly summarized in the final portion of these chapters

Banks are the subject of the second chapter These financial intermediaries are assumed to receive savings and demand deposits from and lend money to households and firms, and to buy government bills and bonds The variables each controls are its own loan rate, the quantity of bills and bonds to purchase, and the maximum amount to be made available for loans during the period. Once these decisions are made, they are presumed to be irreversible for the period in question. Finally, a bank anticipates a rise in the number of its customers when it lowers its loan rate relative to the expected rates of other banks, even though other banks are presumed to follow suit

The representative firm, discussed in chapter 3, "borrows money from banks, hires labor from households, buys goods for investment purposes from other firms, produces goods and sells goods to households and the government The seven main decision variables of a firm are its price, its production, its investment, its wage rate, the amount of money to borrow from the banks, the maximum number of hours that it will pay for, and the maximum number of goods that it will sell" (p 39, symbols omitted) The technology is assumed to be putty-clay As was true for banks, when the firm changes one of its prices (either output price or wage rate), it is assumed that other firms will follow its lead, thereby preventing it from either capturing or losing an ever-greater share of the market

Chapter 4 is devoted to the household sector Each household "receives wage income from firms and the government, purchases goods from firms and pays taxes to the government All goods that are purchased in a period are consumed in that period" (p 75, symbols omitted) At the start of each period, information is received regarding the interest rate on savings deposits, stock prices, the loan rate, the maximum line of credit, cost and (maximum) availability of goods and services, the wage rate, and the maximum recompensable work hours The household's two main control variables are, then, the number of hours of work to provide and the amount to consume

The remaining two sectors, the government and the bond dealer, are considered in the fifth chapter "The government collects taxes from banks, from firms, from households, and from the bond dealer It hires labor from households, buys goods from firms, and issues bills and bonds and pays interest on bills and bonds" (p 97, symbols omitted) The bond dealer, who is assumed to represent the government bill and bond market and the stock market, decides on the bond rate, the bill rate, and the stock price. The interest received on bill and bond holdings and capital gains and losses on bond holdings constitutes the profits of the bond dealer The after-tax income is paid to the household

In chapter 6 the complete set of equations for the "condensed" model is presented, the equations for the "noncondensed" version are found in the appendix. The author points out the links between the financial variables and the real variables and takes care to account for all of the flow-of-funds relationships among the behavioral units by transposing the model in terms of the flow-of-funds accounts As a further test he writes the model in the form of the national income accounts The main part of this chapter (and of the appendix), however, is devoted to a series of computer runs that determine how the economy responds when the values of one or two government parameters are changed temporarily for one period The base run is one in which none of the variables changes from period to period

Whereas chapter 6 and the appendix are devoted to determining some of the dynamic properties of the model, chapter 7 examines a static-equilibrium version of the model and compares the results obtained in the latter with those obtained from the extended IS-LM model Chapter 8 reitereates some of the conclusions, suggests possible extensions, and discusses some of the empirical implications of the theoretical model

I find chapters 1-5 and the conclusion, chapter 8, excellent Although, as with any other model, one could disagree with a number of the underlving assumptions, Fair, to his credit, takes pains to point out where alternative assumptions could have been made and cautions that his results are dependent upon the particular set used He indicates, for example, that the order of information flows and transactions is somewhat arbitrary and adds that a change of order may lead to some differences in the results

My reactions to chapters 6 and 7 and the appendix are mixed Many good points are made in these latter sections, a method for verifying the proper programming of the model and the setting up of the flow-of-funds and national income accounts are two examples I wish, however, that the author had exploited his model more fully The experiments on the dynamic model (in chap 6 and the appendix) involving making temporary changes in the exogenous government policy tools one or two at a time and then observing the impact on the economy Although the author is perhaps correct in implying that a government in many situations may actually be responsible for pushing the economy away from a full-employment equilibrium, it is not clear whether these experiments provide much insight about what the government *should* do in disequilibrium situations More insight could perhaps have been gained had the author run some simulations from a situation of disequilibrium This point is especially relevant when one considers that the model is unstable, this instability makes appropriate government action extremely important Moreover, as a consequence of the nature of the computer runs, many of the results of the simulations are similar to those found in Barro and Grossman (1971), although Fair's model is more complete

Another point worthy of note is found in chapter 7, where the author compares his static-equilibrium version with the extended IS-LM model His static-equilibrium model is a full-employment one, and, therefore, any changes made in exogenous policy tools move the economy to another full-employment equilibrium. The IS-LM model, despite many imperfections, at least makes an attempt to consider unemployment situations

Mv general impression is that the book is an excellent guide for those wishing to learn about macro model building, even though such readers would probably prefer to find the contents of chapters 6 and the appendix switched, that is, to find the "noncondensed" version in chapter 6 and the "condensed" version in the appendix Those who would like to learn about the effectiveness of policy tools in disequilibrium situations may find the book less attractive But, all reservations aside, I do admire Fair's ability to make such a complex model so very understandable

URI M POSSEN

Cornell University

Reference

Barro, R J, and Grossman, H I "A General Disequilibrium Model of Income and Employment" *A E R* 61 no 1 (March 1971) 82–93

An Economic Theory of Learning By RICHARD B MCKENZIE and ROBERT J STAAF Blacksburg Virginia Poly-Tech Institute, University Publications, 1974 Pp 105 $7 95 (cloth), $4 50 (paper)

Economic analyses of the effects of education on personal income and, to some extent, on the individual's consumption have not been scarce in recent years The title and introduction of the McKenzie and Staaf monograph tantalize the reader, suggesting that their study goes beyond the effects of education by examining choices on the types of learning that take place within the educational institution They suggest that they will answer important and interesting questions, such as, "What are the impacts of different student aptitudes and initial endowments of knowledge on student achievement levels and curricular choice?" and "Why is it that four out of ten students switch broadly defined curricula from their freshman to senior year?" Unfortunately, these promises are more easily made than kept, and the reader finds himself increasingly disappointed as he plods through the text

The first substantive chapter begins by setting out a framework in which the student chooses between various loosely defined fields of study according to the way in which they enter his particular utility function The authors go through the standard Lagrangian analysis of consumer choice, where the arguments of the utility function are achievement levels in various academic fields and leisure, and then repeat the analysis in matrix notation The analysis seems to beg the essential question, Why do individuals enjoy academic achievement? The standard analysis suggests that the primary reason rests on wealth augmenting attributes of learning More educated individuals are more productive in the labor market and therefore earn higher wages Simply entering skills into a static utility function ignores this more fundamental issue But even in the context of utility maximization, the analysis is incomplete The only constraint on the individual as modeled is total time and his ability to produce academic achievement. The analysis fails to explicitly consider that the effect of (and perhaps reason for) acquiring skills today is to alter one's opportunity set tomorrow This omission taints the rest of their analysis throughout

The authors go on to point out that knowledge has public-good attributes and use this fact to explain multi-student classes By making the assumption that faculty members derive utility by assigning grades to students in accordance with some monotonic transformation of achievement scores, they derive the proposition that "the ith student's [ordinal] rank not only depends on his own aptitude and effort but on the aptitudes and efforts of the remaining N-1 students" (p 22) An interesting point is that the grading system, by imposing constraints and providing information about ability, is more likely to result in concentration switching among low-ability individuals. One wonders, however, first why performance standards differ among subjects and second, why administrators fail to take this into account when determining satisfactory performance. This leads to the more fundamental question, Why require minimum performances at all? This issue is not addressed by the authors

The final point of the chapter is valid and of some importance. The authors point out that one cannot judge the effect of learning technology simply by examining ex ante and ex post performance in the field in which the technological innovation occurred. Since there is the possibility of using the saved effort to learn other fields, a sufficiently "inferior" field might actually experience poorer performance after the innovation Thus, one cannot conclude that decreasing the student/teacher ratio has no effect even if it does not improve achievement scores

The authors then consider faculty choice and adopt the theme that professors are given sovereignty over what is taught Implications for student and faculty behavior are explored "Publish or perish," they argue, is a rational reaction to the inability to determine what is "good" teaching Although the authors suggest that a voucher system, whereby students could "vote" for good teachers, would reduce the difficulty, they fail to explain the apparent infrequency of such arrangements More generally, the question of what it is that the educational institution produces is not dealt with

The determination of faculty-time allocation is treated as a problem of utility maximization, where the arguments of the utility function are student achievement (a good) and faculty effort (a bad) No unambiguous conclusions are reached, nor is this surprising The utility framework is at best superficial in the case of faculty decisions Later, the choice between teaching and research is considered The faculty utility function is altered to include only teaching rewards and research rewards which, again, remain mysterious The crucial consideration—that the teaching/research decision is likely to affect next period's opportunities—is left unexplored

Chapter 4 discusses "grade inflation " The main point is that formula budgeting, which appropriates revenue to departments according to enrollment, leads to the desire to attract students One method used to accomplish this purpose is to lower the grading standards As all departments attempt to do this simultaneously, grade inflation results This is also applicable across universities, they argue, because the promise of higher grades will induce students to attend that university This point immediately raises three questions First, if grading is a screen, why is it that students learn about the inflation before the users of the screen (e g , their prospective employers)? If the asymmetry is absent, then the students will derive no benefit from the inflation Second, why is grade inflation a recent development? The incentives for grade inflation, if present, should have been present in the past as well If not, why not? Finally, why is the response to formula budgeting higher grades rather than some other student benefit? For example, few would suggest that the response to formula budgeting has been increased faculty time with students

A final substantive chapter considers interdisciplinary study Unfortunately, the authors seems to make a rather basic error by interpreting the *s*-shaped learning curve as implying that knowledge increases at a decreasing rate initially and then at an increasing and finally at a decreasing rate again They confuse the level of the first derivative with the sign of the second, which results in an incorrect analysis This is aggravated by an incorrect statement about the effect of "increasing returns" on optimal investment levels.

The authors go on to argue that interdisciplinary pursuits are more risky as the result of poorer information This, they claim, results in a lower return to interdisciplinary activity Two points should be made First, diversification is usually considered a risk-reducing strategy, so that the argument about relative uncertainty is at least dubious Second, even if uncertainty went in the direction argued, the market equilibrium should be characterized by a higher, not lower, average return to the risky field as long as the marginal investor is risk averse and costs of investment do not differ greatly across pursuits

My general feeling is that although the McKenzie and Staaf book raises some interesting questions, it fails, for the most part, to answer them successfully The book contains many errors, most of them substantive Although the empirical work of others is occasionally cited, no new information is produced The authors frequently derive what they tell us are testable implications but never attempt to test them systematically This is unfortunate, since the theory is insufficiently provocative to stand alone

EDWARD LAZEAR

University of Chicago
and
National Bureau of Economic Research

Wool, Cloth, and Gold The Struggle for Bullion in Anglo-Burgundian Trade, 1340–1478
By JOHN H A MUNRO
Brussels Editions de l'Université de Bruxelles, Toronto University of Toronto Press, 1972 Pp xi + 241 $20 00

In the fifteenth century, a succession of wars and fiscal difficulties stretched the competence of governments in Europe The quarrels between England and France were special instances of these stresses and their inevitable consequences Mounting military expenditures meant a sequence of swelling budgets, and in the search for new ways to augment royal revenues, tax levies dominated domestic issues These problems affected both countries in varying measure France in 1439 underwent a tax reform which came to make the royal taille annual, arbitrary, and permanent England was fortunate enough to avoid arbitrary taxation but only at the cost of shifting the pressure to other sectors amenable to producing a surplus for the king English wool was the best in Europe for certain textiles, and the great cloth-weaving towns of the Low Countries relied on this basic raw material Wool then became a focus of interest. In 1429, as the fortunes in France began to turn, the English monarchy initiated controls on the wool trade as the best way to revive the fisc The staple market for English wool in Calais was put under stringent regulation Merchants were ordered to conform to the "market" price, sell for cash, and deliver a third of the proceeds in money to the royal mint in the city These regulations—the famous Calais Staple and Bullion Ordinances—remained in force for almost half a century, and their effects were both far-reaching and profound. In the end, they served only to accelerate long-term shifts in the wool trade of England, and structural changes in the textile manufactures of the Low Countries themselves

In the economic heyday of medieval Europe, cloth towns such as Ypres had prospered, now they faced crisis and decline. Bruges, which had long remained one of the great entrepôts of northern Europe, also entered a period of acute difficulty, and another port, Antwerp, began to emerge swiftly to take its place

The process of adjustment also saw the rise of a number of new centers concerned with lighter cloths and using wool from Spain. Hondschoote, for example, began to experience a spectacular growth of fortune And in England itself the outcome was dramatic. she soon switched from exporting the raw material to its manufacture Exports of cloth replaced those of wool in a commercial expansion which reached its peak only in the mid-sixteenth century.

John Munro has taken this huge development and produced an interesting book, full of information, which will rightly find its place in the literature. He gives special attention to the arguments on monetary disturbance and regulation, and this is ably supported by evidence from coinage, bullion ratios, and other mint records. The monetary policies of both the Burgundians and the English, he concludes, accentuated fundamental difficulties and eventually resulted in transforming the production and cloth trade of the Low Countries

For good measure in closing the book, he promises to take this research further in another study of textile industries in the north For this, the present work will form a good point of departure

FRANK SPOONER

University of Durham

Growth and Income Distribution Essays in Economic Theory. By LUIGI L PASINETTI Cambridge Cambridge University Press, 1974 Pp x+151 $11 95

Luigi Pasinetti has here collected six essays, three previously published The essays take up Ricardo's system in a mathematical formulation (published first in the *Review of Economic Studies*, 1960), the economics of effective demand, cyclical fluctuations and economic growth (from *l'industria* and *Oxford Economic Papers*, 1960), the historical and analytical transition from classical to twentieth-century "Cambridge school" dynamic analyses, the relation of the rate of profit and rate of growth (from the *Review of Economic Studies*, 1962), and the rate of profit in an expanding economy All except the last essay, as the author tells us in his preface, report work from the period 1958–62, while the last essay dates from 1971–72

These essays are each expressed in a remarkably lucid prose, with accurate mathematical arguments in places where they are appropriate and helpful to the reader Pasinetti's aim is to lay bare the logical structure of the theoretical arguments he discusses, and he succeeds in showing clearly what results flow from what assumptions This is theoretical writing of a very high quality indeed

One of the strongest features of this book is that Pasinetti conceives of "macroeconomics" broadly as the study of all properties of economic systems that are "emergent," that is, that are characteristically apparent only when one looks at an economy as a totality This conception embraces far more questions than those of unemployment, aggregate demand, and inflation Pasinetti argues that distribution, growth, and the rate of profit itself can only be understood as essentially macroeconomic phenomena

There is one lacuna that must be noted at the outset The author's treatment of money and liquidity (macroeconomic phenomena par excellence) is very sketchy and confusing or cryptic in the few places where these topics are treated at all The most distressing problem is Pasinetti's attempt in models that have no explicit representation of money to deduce movements of money prices. For example, on pages 41–42 he presents a Leontief model of production to make the point that the vector of final demands in that system must be chosen so as to be consistent with the available labor He goes on to say that "if [demand] were too high, there would be an inflationary situation " In my opinion this is too much to ask of a model of production that has no representation of money or the formation of money wages and prices The logical leap from a technical incompatibility of final demands and labor input to a movement of money prices ignores too many institutional details to be convincing

When it sticks to an analysis of the movements of real income and its distribution, Pasinetti's discussion is invariably illuminating His paper on cyclical fluctuations and economic growth (the third essay in this book) points out a fundamental problem of macroeconomics multiplier-accelerator models which even at a high level of abstraction can explain a cyclical movement of real income endogenously cannot simultaneously produce steady growth, and vice versa The point he raises here is fundamental even to a wider class of models History concretely has produced a pattern of cyclical expansion of capitalist economies, while economists (with the exception of Marx and a few others who have adopted or modified his formulations) have generally had to be content with attributing either growth (as exogenously given technical progress and increase of population) or cycles (as the result of exogenous shocks to the system), or in some cases both, to unexplained phenomena external to economic theory

This dilemma cries out for a "dialectical" treatment, that is, one which can conceive of quantitative change turning into qualitative change Pasinetti hints at such a resolution in a few paragraphs at the end of his essay He argues that as real income rises consumption patterns constantly shift toward new goods, and this constant shifting will at times confuse entrepreneurs so that parameters such as the speed of adjustment of investment or the accelerator will move from the region where they imply steady growth to the region where the system produces cycles, and back again This ingenious suggestion is not entirely persuasive, since entrepreneurs can influence the evolution of spending patterns through advertising and in fact spend large amounts of resources trying to discover future spending patterns That "mistakes" in this process would add up to a shift of macroeconomic parameters rather than cancelling each other out Pasinetti has not shown An alternative explanation is that the process of capitalist accumulation destroys the preconditions for profitable expansion of production, especially the availability of productive labor, and that this is the qualitative change that leads to an alteration of the parameters of the growth-fluctuations process

Historically the most important and analytically the most powerful of Pasinetti's contributions in this book are his essays on the rate of profit in a growing economy These beautifully reasoned papers attack the formulation that "the rate of profit is determined by the marginal productivity of capital " He assumes that there is a class in society that saves only out of profits (the capitalists) and that the capitalist's propensity to save is larger than that of any other class On a long-run full-employment growth path the capital stock owned by each class will be proportional to its savings, and the profits of each class proportional to its holdings of capital Since capitalists' saving is by assumption proportional to capitalists' profits, this series of proportionalities shows that the growth rate of capitalists' holding of capital, which in the long run is the growth rate of the economy, is equal to capitalists' propensity to save times the rate of profit Since Pasinetti assumes full employment and exogenous exponential growth of the labor force, the economy's growth rate is given, and the rate of profit, regardless of the technology and workers' propensity to save, must be determined by the ratio of the natural rate of growth to capitalists' propensity to save

This proposition is an excellent example of what Pasinetti means by "macroeconomics," since it is an emergent property of an economy that is enforced by systematic changes in techniques and distribution of wealth and income as an economy adapts itself to a long-run steady growth path In this view it is the marginal product of capital (if that concept is well defined) that must conform to the rate of profit through changes in technique.

An intriguing aspect of this controversy, which has been important in developing new results in capital theory and so influential in shaping the thought of many economists, is that the marginalists and the Cambridge economists share many formulations and assumptions. Both study full-employment, steady-state growth paths, both agree that profit-maximizing entrepreneurs must choose techniques so that the value marginal products (when these are well defined) of factors equal factor prices The point is that these assumptions by themselves do not contain enough determinations to close the model, and, as the controversy has made clear, the significance of these shared assumptions changes dramatically depending on the context of other assumptions that are used to close the model

The marginalists, it seems to me, have always had in the "back of their minds" a general intertemporal equilibrium model in which, following Böhm-Bawerk and Irving Fisher, the rate of profit is derived from the relative price of present and future commodities In this system distribution, in the sense of initial ownership of resources, is taken to be determined outside the model, and the model is closed by the *preferences* of households for present and future goods and services These psychological determinations interact with the technological determination to determine the rate of profit Thus if the demand for future goods (in the form of saving) is not sensitive to the rate of profit, it is tempting in this context to say that the rate of profit is determined by the marginal product of capital Of course with many goods and many kinds of capital it is clear that something else besides total demand for future goods (saving) has to be held constant—perhaps the composition of demand or, better yet, the relative prices of future goods But for "pedagogical purposes" a simple case that represents the inherent logic of intertemporal equilibrium can be constructed by taking one kind of capital and one output and suppressing the psychological determinations into an assumption of a constant savings rate

It was on some such slippery slope as this that the Cambridge economists caught the marginalists, and it proved to be much harder for the marginalists to scramble back to firm analytical ground than it was for them to arrive at the one-sector model in the first place Pasinetti and the other Cambridge economists do not begin from a shapeless "general equilibrium" model As Pasinetti says of Keynes and emphasizes also in his discussion of Ricardo, the Cambridge economists believed "that it is one of the tasks of the economic theorist himself (*sic*) also to specify which variables are sufficiently interdependent as to be best represented by simultaneous relations, and which variables exhibit such an overwhelming dependence in one direction as to be best represented by one-way direction relations" (p 44) This methodological posture, together with a willingness to study distribution both of resources and incomes as an endogenous outcome of the growth process, leads Pasinetti to his very powerful critique of the marginalist formulation and to his surprising and dramatic positive results on the determination of the rate of profit Marginalist economists will perhaps always be uncomfortable with and view as "paradoxical" a result that denies to technology any influence, even as part of a simultaneous system, on the rate of profit. Pasinetti's essays here, however, make a persuasive case that the "marginal productivity of capital," even when it is a well-defined concept, is not except in special cases an important determinant of the rate of profit on a long-run full-employment growth path It would appear that the influence of technology on the rate of profit can only be sustained by banishing distribution of resources from the theory and reintroducing subjective psychological determinations, as general equilibrium theory does These results of Pasinetti's building on other work of Cambridge economists constitute, in my view, a very great contribution to economic theory

But perhaps we need to carry the methodological moral of this tale even further *If the context of assumptions turns out to be vital to their significance we ought* to be extremely critical of every assumption in these highly abstract models While Pasinetti is extremely acute in dissecting the marginalists' treatment of technology and saving, he is somewhat complacent about the assumption of full employment of labor While this is a persuasive assumption for particular national economies with highly developed capitalist production, it is not so persuasive if we look at capitalism as a world system, since on a world scale capitalist production appears to operate with a large margin of unemployed and underemployed labor A revision of Pasinetti's assumption of full employment of an exogenously given labor force would entail considerable further dramatic changes in the theory of distribution and growth. As Pasinetti writes "The worst of Malthus' and Ricardo's fears, and the most frightening of Marx's predictions (that real wage rates would be nailed down to subsistence and all surplus above subsistence would be appropriated by the capitalists) turn out to be incompatible with the maintenance of full employment" (p 102) But whether this observation should lead us to reject Ricardo's fears and Marx's predictions or the assumption of full-employment growth paths with a labor force given outside the accumulation process remains an unsettled question

Beyond Pasinetti's striking results in growth and distribution theory, I think the methodological hints and examples of this book are extremely important. Pasinetti's emphasis on emergent properties of economic systems, his support of one-way (which he somewhat apologetically and inaccurately calls causal) determinations, his aim of achieving sharp results from simple, strong, assumptions, and his sensitivity to the importance of the axiomatic context for particular ideas (what might be called the gestalt of economic models) are all exciting, and to my mind valuable, corrections to the growing body of economic theory that is formally elaborate but empty of substance

DUNCAN K FOLEY

Stanford University

Lancashire in Decline A Study in Entrepreneurship, Technology, and International Trade

By LARS G SANDBERG

Columbus Ohio State University Press, 1974 Pp xii +276 $15 00

A reader of this *Journal* does not need to be told that in the last decade or so there has arisen a new brand of economic history, with the emphasis on "economic," which has sharply revised the story of economic development in America. He may not be aware that recently the "new" economic history has moved on to other stories, in particular to the story of America's nineteenth-century predecessor as Top Nation—Britain. Lars Sandberg's book, a distinguished contribution to this growing stream, is directed at the widely held premise in economic history that old industrial countries, Britain then as America now, experience entrepreneurial failure. Since the last quarter of the nineteenth century, when Englishmen awoke to find Germans and Americans overrunning their traditional markets abroad, alleged evidence for the premise has been mounting, and by now it is taken for granted that even in Britain's characteristic contribution to industrialism —cotton mills—British managers dallied in adopting new techniques and were therefore crushed in old markets

Sandberg's argument is a very different one, and more persuasive. He argues in the first half of his book that the chief examples of entrepreneurial failure in British cotton textiles, notably the slow move from the mule to the ring spindle, are poorly chosen, for the techniques Englishmen adopted were in fact rational in view of the factor prices they faced. He argues in the second half that the natural spread of industrialization and of tariffs to protect it, rather than entrepreneurial failure at home, explains Britain's falling share of the world's market for cotton textiles

In common with other pieces of new economic history, the reasoning used is often startlingly simple, although it takes a sophisticated economist to cast the issues into a simple form Sandberg is able to show that the returns to the new technologies of the ring spindle and the automatic loom were lower in Britain than in America, for example, simply by recasting the evidence into the form of a cost-benefit analysis, or that larger investment in new technologies before World War I would have resulted in larger output but also lower profits after the War simply by noting that world prices would have fallen even further, or that Britain's position of comparative advantage in exporting cotton textiles was fragile and finally broken by Japan simply by emphasizing the low costs of transporting finished cloth The weakness of the old economic history was not usually in collecting facts (although Sandberg develops some new statistical series, notably an index of the quality of British textile exports, on the whole he builds on the work of earlier scholars) but in refining the facts that matter and interpreting them in the light of simple economic theory The theory in the book, indeed, is occasionally oversimple It does not really follow from Britain's dominance of Indian imports of cotton textiles before World War I, for example, that Britain needed to bear the brunt of the 60-percent fall in those imports after the War (p 182), for the market was a world one Here and elsewhere in the book, especially in its second part, the arithmetic of market shares sometimes dominates the economics of market supply and demand curves Yet the economics is seldom far below the surface of the argument, and the argument is always lucidly expressed.

If it is the applied economics in the new economic history which distinguishes it from most economic history, it is the history in it which distinguishes it from most applied economics It is the historian in Sandberg who demands that he devote half his book to gathering evidence on the relevance of *homo economicus* to the British textile industry in the late nineteenth century rather than, as the economist in him would prefer (unless he were an agricultural economist disputing the idiot peasant model in India or Nigeria), simply assuming it relevant at the outset It is the historian who demands he devote the other half to applying the model so validated to half a century's history rather than to last year's. And, finally, it is the historian who demands he write the book on Britain rather than America, for America's economic experience has been rich, pacific, and democratic—that is to say, from the perspective of Russia or India or Britain, strange in every way The economic history of America is, certainly, of great interest, but more of parochial than of cosmopolitan interest If economies other than the

American are to be understood, their history must be studied, and studied with the tools of modern economics Indeed, Sandberg's work, part of a growing body of work on Britain tending to the same conclusion, suggests that Britain's economic history in the last third of the nineteenth century is more applicable to America's in the last third of the twentieth in America now, as in Britain then, the news of her industrial death has been greatly exaggerated

DONALD N MCCLOSKEY

University of Chicago

Market Signaling Informational Transfer in Hiring and Related Screening Processes
By A MICHAEL SPENCE
Cambridge, Mass Harvard University Press, 1974 Pp 221 $10 00

Today, the tuition per student per lecture at a big private university is around $10 This is rather humbling, how many lectures do we give that are worth that much? Michael Spence in this lucid and interesting book gives us good cheer We may be providing useful information, but about our students rather than to them

His basic idea is that, since hiring and firing are expensive, employers (or anyone buying services from unestablished individuals) would like to know in advance how productive their employees will be There are two types of information they can use to judge applicants unalterable characteristics such as race and sex and characteristics such as education which can be manipulated by the applicant at some cost to him Spence calls the latter "market signals " Provided that the costs of employee signals are sufficiently negatively correlated with productivity, employers can use them to select more productive people

This is a call to us as teachers Suppose firms seeking intelligent drudges make the B A a prerequisite to well-paying jobs If we can make our classes unbearable to the other potential job seekers, only the drudges will get the B A and the jobs The situation will be in "informational equilibrium," because the employer's expectations about what the requirement meant would be fulfilled Spence makes the interesting points that education may be unproductive in itself and remain useful as a signal, that multiple equilibria are possible (for example, firms might equally well require an M A and get essentially the same people, burdening them with the extra schooling), that equilibria persist like putty at whatever requirements currently prevail, and that since disconfirming evidence may never appear firms may maintain different requirements for identifiably different groups such as blacks and whites, even though their economic potential is identical

Spence shows that if firms compete on signal requirements all firms will use the least-cost discriminating signal and so multiple equilibria will not exist He implies that such competition is unlikely since the costs of unnecessary education are borne by individuals and not firms, and the firms would therefore have little incentive to compete by reducing requirements This last assertion seems wrong, in exchange for labor services firms provide both wages and a job The attributes of the job including its requisite signals are tied to the wages, and equality of marginal product with wages must be interpreted in this wider sense Thus tuition may be subtracted when there is learning on the job, and in a market where useless M A 's are the standard a firm can offer a B A -and-lower-wages job and compete effectively for the best new people Indeed, a better ability to discriminate is an asset that casting and employment agencies rely on for their profitability.

The theory has something to offer in understanding many other economic phenomena, including guarantees, brand names, internal promotion, certifying agencies, and discrimination In discussing the case of DeFunis (the student who sued the University of Washington Law School for admission on the grounds that they had accepted black students with lower grades and test scores than his), Spence points out that if the costs of educational signals are higher for blacks, blacks with the same productivity as whites may have less impressive academic records In this case discrimination is unavoidable at some level—signals will be different if people are chosen by productivity, or productivity will be different if people are chosen by signals (as DeFunis argued they should be) Most people would probably prefer to use future productivity rather than test scores as the criterion of "merit " Unfortunately, there is little evidence of how future productivity varies by group with test scores, and it might be hard even to agree on a good measure of postschool productivity Thus even if the assumption Spence would use to justify different test standards for different groups were true, it would be difficult to validate

We could make a quite different argument than the one Spence employs against the use of uncompensated test scores This argument does not rely on the differential signaling cost assumption In such cases as professional school entry, union membership, or government jobs, the state (by cash or statute) may be subsidizing positions that have more qualified applicants than vacancies This subsidy may make "division of the spoils" a more appropriate selection principle than merit, however defined Every group with political power should get its fair share

The author presents his many new insights by simple examples, neat arguments, and clear concise prose Reading this book is a joy

EMMETT KEELER

RAND Corporation

Late Nineteenth Century American Development A General Equilibrium History By JEFFREY G WILLIAMSON
London and New York Cambridge University Press, 1974 Pp 350 $19 50

In this book Jeffrey Williamson develops a 72-equation, two-region, three-sector, three-factor neoclassical growth model, intended to represent the American economy between 1870 and 1910 Equipped with parameter values obtained through a melange of ad hoc procedures and rough guesses, the model generates 40 hypothetical years of history, telling an economic story which, in the author's view, remarkably resembles the real economic history of this period Having received high marks from its creator on its ability to reproduce the actual record, the model moves on to tell a variety of counterfactual tales of the same period, histories which "would have been" had this or that factor been different in this or that way Among the more surprising of the conclusions are these (1) the major features of American growth are not associated with the closing of the frontier or the rise of financial intermediaries but may be explained most directly by the postwar acceleration in the rate of capital formation and all its ramifications, (2) the fall in interregional transport costs, however, had a major effect on resource allocation and growth, (3) industrialization was primarily a result of the exogenous effect of world market conditions, (4) farmers were fully justified in complaining of "capital starvation" in the early 1870s, but later in the period indicators of farm distress are more than offset by the capital gains on farmers' landholdings

Despite its size, the structure of the model is essentially quite simple. Agricultural goods are produced in the Midwest by land, labor, and capital Industrial commodities are produced in both the Midwest and Northeast by capital and labor only Production functions are Cobb-Douglas in all sectors. Commodity prices are established by world markets and are unaffected by developments in America Factors are paid their marginal products, however, labor and capital markets do not equilibrate instantaneously Instead, migration between sectors and regions is a function of differentials in wages and interest rates In the "basic" simulation of chapter 4, aggregate labor-force growth and the savings rate are exogenous, although this feature is partially relaxed in later chapters The growth of the land stock, labor-force participation, transportation costs, and total factor productivity are exogenous throughout

This is a challenging book, and it raises some basic questions about our understanding of the historical process of economic development Does a highly aggregated, full-employment, strictly nonmonetary neoclassical growth model—albeit one with significant disequilibrium in capital and labor markets—really capture the essence of American economic experience during this period? I am far from convinced Should we expect such a "long-run" model to give a close fit over a 3-decade period marked by drastic monetary and institutional changes, a period containing 2 separate decades in which (according to Lebergott) unemployment *averaged* 10 percent of the labor force (not "approaching 10%," as Williamson has it on p 117)?[1] Williamson fully acknowledges (pp 70, 78, 84, 212) that the model does not satisfactorily reflect business cycles or long swings in economic growth, but he does not explain the osmotic process by which these cyclical mechanisms generate long-run full-employment growth results in historically short periods of time From this perspective, is it really playing this simulation game fairly to apply the analysis to one period which seems to fit a neoclassical outline on a prima facie basis (as Temin and others have noted), and then to terminate the experiment when the model stops working—with rather vague reference to "structural change," apparently unconnected with identifiable historical developments, economic or other? Williamson indicates (pp 5, 118) that he views the "epochal hinge" after 1896 as a major unsolved problem, but he does not suggest that it might be grounds for doubting his analysis of the earlier years Indeed, if one is willing to allow the existence of long swings, as Williamson apparently is, can one even *define* "secular" trends with any confidence over so short a period? The long-swing hypothesis, for all its conceptual vagueness, at least has the virtue of explaining both the retardation of the 1890s and the acceleration after 1900 in a single package, rather than regarding the first as "solved," the second as "left for future research "[2]

---

[1] I say "3-decade" even though the actual simulation covers 40 years, because at several points (e g , pp 5, 31, 118) it is conceded that 'structural change" after 1900 renders the model inappropriate for some or all of the last decade While he thus disclaims responsibility for these years, Williamson is nonetheless unable to restrain himself from claiming credit (on behalf of his model) for those post-1900 developments which the model does predict (e g , pp 75, 83, 130-31) Heads I win, tails you lose The attitude toward long swings is similar

[2] I am unfortunately stubborn enough to be still stuck on an earlier point just how much "retardation" occurred, apart from the worldwide industrial depression of the 1890s I am not persuaded by the claim that the depression of the 1870s was "equally severe" (p 93) in terms of the loss of output It requires some fancy footwork with quadratic time trends (appendix C) to generate substantial industrial "excess capacity" in the 1870s, a finding which I find grossly implausible on the basis of the raw data

The reader who is skeptical, as I am, will find this a frustrating book, because Williamson does not elaborate or defend in a serious way the outlook toward economic history embedded in his model. Instead, the book relies heavily on the ability of the model to "track" the major historical developments of the period However, I do not find this performance as impressive as does Williamson True enough, the model is shown to reproduce many trends in aggregate GNP and capital-stock growth (pp. 68–74), but just what does this mean? In the basic simulation of chapter 4, it is only a slight exaggeration to say that Williamson has "explained" the growth of output in terms of three exogenous factors of production and an exogenous "residual," which was itself estimated by dividing output by an index of inputs Land-stock growth is fully exogenous, and labor-force growth is virtually exogenous, capital-stock growth is technically endogenous, but the savings rate is exogeneous—and indeed, since the initial capital stock and the depreciation parameter are also exogenous, the "savings rate" estimates employed are simply another way of expressing the same information contained in the (actual) capital-stock figures Of course the fit is good As far as I can see, the only reason the fit is not perfect is that Williamson has slapped together estimates from various scholars using different weighting schemes and conceptual specifications, and because for many variables he obtains 40 "annual" observations by interpolation between the *five* census-year data points actually available How can this procedure possibly tell us whether the model is "fact or fiction"?

How much of the tracking of various economic magnitudes is strictly tautological I am frankly uncertain, but Williamson gives us no help in separating the wheat from the chaff I find no mention of the risks of pseudoverification, no inquiry into the informational sources behind the data themselves Even where the "actual" observations seem relatively independent of the parameter estimates —as with the real wage figures (pp 78–83)—I am dubious Not only was the labor elasticity obtained from a factor share estimate, but the book indicates that the main determinant of regional wage differences was the differential in commodity prices With this in mind, I am unable to work up enthusiasm for a model which "predicts" real wages by assuming all commodity prices to be exogenous None of this proves that the Williamson model is wrong, but I do not find compelling reasons here for believers in alternative models to change their minds Their persuasion will not be furthered (quite the contrary) by demonstration that the model's counterfactual implications are surprising and "revisionist "

I do not mean to imply that all of the hypotheses of the book rise or fall with the success of the simulation exercise To a large extent, readers can simply be advised to take chapter 4 with a grain of salt and move on to the substantive chapters, which offer a diverse menu of cliometric fare, chewable if not always palatable or digestible But, unfortunately, since the purported accuracy of the general simulation is apparently the reason for the cursory and frequently casual defense of basic assumptions, the hypnotic effects of the computer pervade the whole book For example Williamson does note that the main reason his assessment of the role of land and agriculture is so different from previous writers is that he assumes "small-country" conditions, treating all commodity prices as exogenous It does not take a 72-equation computer model to prove that the "spector of diminishing returns" can be banished by this assumption.[3] Chapter 10 on "World Markets" shows that the simulation results are highly sensitive to hypothesized changes in these relative prices. One would have thought, therefore, that establishing the appropriateness of the exogeneity assumption would be a major concern of the book. (Quite clearly the success of the general simulation has no bearing on the point: one can hardly "predict" prices any better than by assuming they had to be what they were.) But instead we find only one brief and (in my view) unsatisfactory discussion (pp. 217–20), limited to wheat in the British market.[4] Surely we need more than this to sweep away with one stroke of the broom traditional historical themes relating to the "rise of a national market," involving *changes* in the effective demand elasticities facing producers and major alterations in patterns of specialization within agriculture.

---

[3] Note also, however, that industry uses no lumbe., coal, oil, iron ore, or any other natural resource in this model

Chapter 6 on financial intermediaries develops an interesting alternative to the "institutional" explanations for the pace of regional interest-rate convergence between 1870 and 1900, but here again I find that the critical assumptions go virtually undefended. Williamson's "disequilibrium" explanation emerges from his "putty-clay" assumption that new savings are required to effect capital migration. The flow of capital could not keep up with the very rapid westward expansion of labor during the 1870s; interest rates converged only when the movement of the "frontier" slowed down. But this result is only possible because the model, which treats the economy as "open" in terms of commodities, assumes a "closed" economy when it comes to capital. The justification offered (pp. 33, 309–10) is that net foreign capital imports were not in fact large when averaged over the whole period, though they frequently were during "peak" years. (Query: just what years *does* the full-employment model apply to if not the peaks?) The problem with this argument is that *gross* capital flows were often quite large, suggesting that a worldwide shortage of "new savings" was not really the constraint on interregional capital flows. Granted that *net* flows were not large over the period as a whole, the question remains: why not? Williamson begs the question, avoiding "institutional" considerations only by means of his specification.[5] Even taken on its own terms, the argument would seem less plausible if the South were included. If the interest-rate differentials were opened up by the pace of *demand*, why should the slow-growing South have suffered under even higher interest rates, which fell even more slowly, than those of the Midwest?

Here and elsewhere, the ground rules are always the same: if Williamson can generate an observed pattern with his model, then there is no "need to rely" on other arguments; unfortunately, this often seems to mean neglect of the analytical and historical dimensions of the subject. As a result we get very little grasp of just what kind of approximations we are dealing with and how the simulations relate to other types of knowledge and information. I am sure that the many specialists whose work falls under criticism in the book will feel that they have not received a reasonable reading or refutation from Williamson, and I am afraid that many will feel that he does not even warrant a response. This would be a pity, because the book is a major research effort by a competent scholar, and it is clearly intended as an appeal to economic historians to rethink their positions on a variety of issues and to propose and discuss their alternative hypotheses within this or a similar framework. Rethinking is of course always in order, and many of the particulars are well worth the attention of scholars, but I do not share the hope that the approach of the book will be widely imitated My own undoubtedly biased opinion is that we still have a great deal to learn about basic behavioral parameters in economic history and that the choices in specification are very great indeed High-level computer simulations do not seem to be a good substitute for the direct confrontation of these issues

---

[4] An earlier section is labeled "a defense of the 'small country' assumption," but it contains no specific defense. At one point (pp. 216–17) Williamson even seeks to bolster his case on industrialization and world markets by reference to *exports* of "heavy industry," offering as an explanation only the note that "imports of manufactured products are to be understood as imports net of exports of such products."

[5] I do not comprehend the statement (p. 129) that the model does not "minimize the importance of institutional change." How can the nonreversible developments in the financial sector be embedded in the (stable) parameters of the savings-transfer function if they do not effect the "equilibrium" differential?

GAVIN WRIGHT

University of Michigan

CHICAGO

THE CITIZEN AND THE STATE
Essays on Regulation
George J. Stigler

Stigler writes with intelligence and humor, and his stimulating essays should encourage a reexamination of the assumptions about what government regulation of economic affairs can and does accomplish *The Citizen and the State* will be of great interest to the business community and to students of politics and economics
1975 224 pages Cloth $10 95

ENERGY: THE POLICY ISSUES
Edited and with a Foreword by Gary D. Eppen
With a Preface by Harold S Geneen

Eminent scholars from a number of fields discuss the major themes of American energy policy Among the topics covered are federal regulation of petroleum prices, the natural gas industry, and ecological abuse in energy production Published under the auspices of *The Journal of Business*
1975 Cloth $9 50 September

Now in Paper·

THE SUPREME COURT AND PATENTS AND MONOPOLIES
Edited by Philip B Kurland
1975 Paper TBA

THE UNIVERSITY OF CHICAGO PRESS
Chicago 60637

A Classical Macroeconometric Model for the United States

Thomas J. Sargent

University of Minnesota

A statistical definition of the natural unemployment rate hypothesis is advanced and tested. A particular illustrative structural macroeconomic model satisfying the definition is set forth and estimated. The model has "classical" policy implications, implying a number of neutrality propositions asserting the invariance of the conditional means of real variables with respect to the feedback rule for the money supply. The aim is to test how emphatically the data reject a model incorporating rather severe classical hypotheses.

This paper estimates a small, linear, classical macroeconometric model for the postwar United States. One reason for estimating the model is to produce a simple device capable of generating unconditional forecasts of key economic aggregates such as the unemployment rate, the price level, and the interest rate. But a more important reason is that as part of the estimation process the hypotheses underlying the model are subjected to empirical testing. Since these hypotheses are very "classical" and sharply at variance with Keynesian macroeconomics, it would be useful to know at what confidence levels the data reject them.

The present model is considerably more monetarist than is the St. Louis model.[1] Indeed, as interpreted and manipulated by its builders, the St. Louis model is incapable of rationalizing prominent monetarist positions. In particular, it implies that simple *x* percent growth rules for money can generally be improved upon by adopting rules with feedback from past endogenous variables to current money.[2] By way of contrast,

---

This paper was financed by the Federal Reserve Bank of Minneapolis, which does not necessarily endorse the opinions expressed. Thomas Turner, Paul Anderson, and Salih Neftci performed the calculations.

[1] See Andersen and Carlson 1970.

[2] Cooper and Fischer (1974) have made this point.

[*Journal of Political Economy*, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved.

the present model is one in which an x percent growth rule for the money supply seems not to be dominated by any rule with feedback.[3]

The deterministic (nonrandom) classical model, the static analysis of which is enshrined in macroeconomics textbooks, has never been taken seriously, because its predictions seem so terribly at variance with the data In particular, it is hard to explain the observed persistent movements in employment and unemployment with the textbook classical model. How meaningfully integrating random, disturbances into the classical model would affect the analysis is a matter about which there is presently little agreement On the one hand, in his American Economic Association presidential address, James Tobin (1972) seemed to assert that the presence of random disturbances in demand and supply schedules so alters the character of the general system that it sets up an exploitable trade-off between unemployment and inflation even in a system where all agents optimize On the other hand, Robert Lucas (1972*b*) has analyzed a general equilibrium system in which agents cope optimally with the existence of uncertainty While there exist "nonneutralities" in that system, there are no nonneutralities that the government can either exploit or offset by way of countercyclical policy

This paper formulates, tests, and estimates a version of the classical model that has its origin in hypotheses that place severe restrictions on the random behavior of unemployment, output, and the interest rate The model implies that those three "real" variables are econometrically exogenous with respect to variables measuring monetary and fiscal policies As a consequence, government manipulations of monetary and fiscal policy variables have no predictable effects on unemployment, output, or the interest rate and hence are useless for pursuing countercyclical policy Such implications are in the nature of neutrality results, albeit ones that require drawing some fairly fine econometric distinctions The key elements of the model that provide the sources of the restrictions on the stochastic nature of output, unemployment, and interest are (*a*) a drastic version of the natural unemployment rate hypothesis, (*b*) the expectations theory of the term structure of interest rates, and (*c*) the assumption that the public's expectations are "rational "

The chief novelty of this paper is its formulation of a drastic, statistical definition of the natural unemployment rate hypothesis That definition is not dependent on any particular macroeconomic structural model, being compatible with a variety of structures one could imagine The particular structural model presented in this paper is intended only as an illustrative example that satisfies this definition of the natural-rate hypothesis This particular structure does, however, illustrate some of the strong classical properties that will be possessed by models that satisfy the definition of the natural-rate hypothesis advanced here A major aim of the paper is to indicate how this definition of the natural-rate hypothesis can be tested and to present some test results

---

[3] Models with this property have previously been analyzed by Sargent (1973) and Sargent and Wallace (1975)

This paper is organized as follows In section I a prototype of the model is described However, no attempt is made here to rationalize in a deep way the equations comprising the model Section I is designed to display the system briefly and to establish its classical nature. Section II then provides a definition of the natural-rate hypothesis that is the cornerstone of the model Statistical tests of the hypothesis are described. Section II also describes how the rational-expectations theory of the term structure of interest rates is implemented in the model and how its central implications can be tested Section III implements the econometric tests described in Section II Finally, Section IV contains estimates of the complete model The casual reader not interested in econometrics can read only Sections I and IV and find there estimates of the model and a description of how it works

I. Overview of the Model

I begin by describing a simple prototype of the model It differs from the model finally estimated in some minor ways but illustrates well the mechanics of the model

The prototype consists of the following five equations

$$Un_t = \gamma(p_t - E_{t-1}p_t) + \sum_{i=1}^{n_1} \lambda_i Un_{t-i} + u_{1t}, \tag{1.1}$$

where $\gamma < 0$ (a Phillips curve),

$$n_{ft} = \beta(p_t - E_{t-1}p_t) + dUn_t + \sum_{i=1}^{n_2} w_i n_{ft-i} + u_{2t}, \tag{1.2}$$

where $\beta > 0$, and $d < 0$ (a labor force participation equation),

$$y_t = \alpha_0 t + \alpha_1(n_{ft} - Un_t + pop_t) + u_{3t}, \tag{1.3}$$

where $\alpha_1 \simeq 1$ (a production function);

$$R_t = R_{t-1} + \xi(Z_t - E_{t-1}Z_t) + u_{4t} \tag{1.4}$$

(a martingale equation for the long-term interest rate); and

$$m_t - p_t = b_1 R_t + b_2 y_t + b_3(m_{t-1} - p_{t-1}) + u_{5t}, \tag{1.5}$$

where $b_1 < 0$ and $b_2, b_3 > 0$ (a portfolio-balance schedule). The variables are defined as Un_t = unemployment rate, p_t = log of GNP deflator;

n_{ft} = log of labor force participation rate; y_t = log of real GNP, pop_t = log of population; R_t = long-term interest rate; m_t = log of money supply, Z_t = a vector of exogenous variables in the "IS" curve, including tax rates and government purchases, u_{jt} = mutually and serially independent random terms with zero means, so that $E_{t-1}u_{jt} = 0$, $j = 1, \quad , 5$, and $E_{t-1}X_t$ = the mathematical expectation of X_t conditioned on information available at time $t - 1$ The variables Z_t, pop_t, and m_t are taken as exogenous.

Equation (1 1) is a Phillips curve that posits an inverse supply-side relationship between unemployment and the unexpected part of the current price level The public's psychological expectation about the price level is supposed to be "rational," meaning that it equals $E_{t-1}p_t$. The equation embodies the natural unemployment rate hypothesis, since it asserts that unemployment does not depend on the anticipated part of the rate of inflation Equation (1 1) is essentially Lucas's (1973) formulation of the Phillips curve

Equation (1 2) is a labor force participation equation positing that the participation rate depends directly on the unexpected part of the price level and inversely on the unemployment rate (the "discouraged worker effect") The presence of unemployment and the unexpected part of prices in equations describing labor force participation is not unusual (e g , see Wachter [1972] and the work cited by him)

Upon noting that the log of employment approximately equals $(n_{ft} - Un_t + pop_t)$, equation (1 3) is seen to be a Cobb-Douglas production function that excludes capital The regressions reported by Bodkin and Klein (1967) and Lucas (1970) suggest that little violence is done to the data by omitting capital from equation (1 3) That is, time-series regressions of the log of output against the logs of capital and employment typically display constant or increasing returns to employment and zero or slightly negative returns to capital For my purposes, excluding capital from equation (1 3) permits the construction of a model in which there is no need to account for capital accumulation

Equation (1 4) posits that the long-term interest rate is a martingale Fiscal policy and other aggregate-demand variables influence the long rate in two ways First, the unexpected components of Z_t influence the "innovation" in the long rate, that is, the part of the long rate that cannot be predicted from the past Second, the foreseen or expected part of Z_t is already reflected in R_{t-1} and affects R_t in precisely the same way it affects R_{t-1}

Equation (1 5) is a standard portfolio-balance schedule.

The model is five equations in the five endogenous variables Un_t, n_{ft}, Y_t, R_t, and p_t The exogenous variables are Z_t, pop_t, and m_t

To complete the model, the stochastic processes governing the exogenous variables m_t, Z_t, and pop_t must be specified. I will assume the autoregressive schemes

$$m_t = \sum_{i=1}^{n_3} \xi_i m_{t-i} + \varepsilon_{1t}, \tag{1.6a}$$

$$Z_t = \sum_{i=1}^{n_4} \psi_i Z_{t-i} + \varepsilon_{2t}, \tag{1.6b}$$

and

$$pop_t = \sum_{i=1}^{n_5} \omega_i pop_{t-i} + \varepsilon_{3t}, \tag{1.6c}$$

where the ξ_i's, ψ_i's, and ω_i's are parameters, and the ε_t's are serially independent random variables with means of zero; they are assumed to be distributed independently of the u's in the structural equations (1.1)–(1.5). To solve the model and to forecast with it, expected values of the exogenous variables, for example, $E_{t-1}m_t$ and $E_t m_{t+1}$, are required. These expected values are calculated using the autoregressions above for the exogenous variables. This is partly by way of imposing rationality, since the expected price $E_{t-1}p_t$ turns out to depend on $E_{t-1}m_t$, $E_{t-1}Z_t$, and $E_{t-1}pop_t$. Rationality amounts to requiring that the public's expectations of the exogenous variables m_t, Z_t, and pop_t equal the mathematical expectations computed from the appropriate objective probability distributions, that is, the autoregressions above.

The model has a standard aggregate-demand and -supply representation in the p, y plane. Substituting equations (1.1) and (1.2) into (1.3) gives the aggregate-supply schedule

$$\begin{aligned} y_t = {} & \alpha_0 t + \alpha_1 pop_t + [\alpha_1 \beta + (\alpha_1 d - \alpha_1)\gamma](p_t - E_{t-1}p_t) \\ & + (\alpha_1 d - \alpha_1) \sum_{i=1}^{n_1} \lambda_i Un_{t-i} + \alpha_1 \sum_{i=1}^{n_2} w_i n_{f\,t-i} \\ & + (\alpha_1 d - \alpha_1)u_{1t} + \alpha_1 u_{2t} + u_{3t}. \end{aligned}$$

Since $\alpha_1 \beta_1 + (\alpha_1 d - \alpha_1)\gamma > 0$, the aggregate-supply schedule is upward sloping in the p–y plane.

Substituting equation (1.4) into (1.5) gives the aggregate-demand schedule $p_t = m_t - b_1 R_{t-1} - b_1 \xi (Z_t - E_{t-1}Z_t) - b_2 y_t - b_3(m_{t-1} - p_{t-1}) - b_1 u_{4t} - u_{5t}$, which slopes downward in the p, y plane. Increases in m_t and in the aggregate-demand innovations $\xi(Z_t - E_{t-1}Z_t)$ cause the demand schedule to shift outward. The equilibrium p, y combination is determined at the intersection of the demand and supply curves.

While the model is clearly simultaneous in determining the current values of the five endogenous variables, for generating forecasts it is recursive. The one-period-ahead forecast of Un_t is determined by taking expectations in equation (1.1) conditional on data known at $t - 1$:

$$E_{t-1}Un_t = \sum_{i=1}^{n_1} \lambda_i Un_{t-i}, \qquad (1.1')$$

which follows since $E_{t-1}(p_t - E_{t-1}p_t) = E_{t-1}p_t - E_{t-1}p_t = 0$. The forecast of n_{f_t} is then given from equation (1.2) as

$$E_{t-1}n_{f_t} = dE_{t-1}Un_t + \sum_{i=1}^{n_2} w_i n_{f_{t-i}}$$

Then from equation (1.3) we have the forecast of the log of GNP as $E_{t-1}y_t = \alpha_0 t + \alpha_1(E_{t-1}n_{f_t} - E_{t-1}Un_t + E_{t-1}pop_t)$. From equation (1.4) the forecast of the long-term interest rate is simply $E_{t-1}R_t = R_{t-1}$, which follows since $E_{t-1}(Z_t - E_{t-1}Z_t) = 0$. Finally, from the portfolio-balance schedule, the forecast of p_t is $E_{t-1}p_t = E_{t-1}m_t - b_1E_{t-1}R_t - b_2E_{t-1}y_t - b_3(m_{t-1} - p_{t-1})$. To compute the forecasts of the endogenous variables, the forecasts $E_{t-1}m_t$ and $E_{t-1}pop_t$ of the exogenous variables are required.

The predictions of the model are obviously classical in spirit. The predictions of the "real" variables are all independent of the prediction of the money supply, which only influences the predicted price level. For predicting the long-term interest rate, predictions of the fiscal and other aggregate-demand variables add no information to that in the current long rate since they are already properly embedded in the current long-term rate. Finally, the model implies that the monetary authority does not have the option of pegging the nominal interest rate R_t via some feedback rule by letting the money supply be whatever it must to guarantee portfolio balance at that interest rate.[4] For suppose that the authority were to attempt to peg the interest rate via the feedback rule

$$R_t = F\theta_{t-1}, \qquad (1.7)$$

where θ_{t-1} is a vector of observations on endogenous and exogenous variables dated $t - 1$ and earlier, and F is a vector of parameters conformable with θ_{t-1}. The predictions of R_t from equations (1.4) and (1.7) are clearly in general inconsistent, so that the interest rate is overdetermined. Thus, this model is characterized by Wicksell's classical overdeterminacy of the interest rate (and indeterminacy of the price level) under a pegged interest rate.

It bears emphasizing that, while for prediction the model has a very classical recursive structure, it is a simultaneous model when it comes to determining current variables. Thus, money is not a "veil" in the model, since (random) increases in money can be shown to stimulate both GNP and the price level. So will (random) increases in the aggregate-demand Z's But it turns out that in this model it is best to predict as if money were a veil The fact that variables are determined jointly simply cannot be exploited in prediction, neither can it be exploited for control

---

[4] This is one of the options analyzed for a stochastic Keynesian model by William Poole (1970).

I have indicated that to generate forecasts of the endogenous variables the exogenous variables should be set equal to the forecasts $E_t m_{t+1}$, $E_t Z_{t+1}$, and $E_t pop_{t+1}$, which are to be computed from the autoregressions (eqq [1 6a]–[1.6c]) that actually govern those exogenous variables It seems that something more is possible in the way of forecasting, but it turns out not to be useful to the policymaker In particular it is possible to use the model to "predict" values of the endogenous variables in $t + 1$, conditional on alternative assumed values for the exogenous variables m_{t+1}, Z_{t+1}, and pop_{t+1}, *given* values of $E_t m_{t+1}$, $E_t Z_{t+1}$, and $E_t pop_{t+1}$ For example, for a given $E_t m_{t+1}$, different values of m_{t+1} will be associated with different values of the *real* variables output and unemployment The larger $m_{t+1} - E_t m_{t+1}$ is, the larger will be the "predicted" value of output and the lower the "predicted" value of unemployment But such "conditional" forecasts are of no use in forming policy For example, it will not work to use the model to "forecast" unemployment for alternative values of m_{t+1}, given $E_t m_{t+1}$, and then to set m_{t+1} in order to achieve the unemployment rate desired by the monetary authority Expecting that to work amounts to assuming that the public would continue to form its expectations about m_{t+1} by using equation (1 6a) even if the authority adopted the new and different rule for setting m implicit in the procedure above That violates the assumption that expectations are rational What affects unemployment and output is the gap between m_{t+1} and $E_t m_{t+1}$, and there is no way that the authority can expect to set this gap at some desired nonzero level

This completes the overview of the model I now turn to the task of setting forth more precisely the nature of the key hypotheses underlying the model In the process, statistical tests of those hypotheses will be described and implemented

II. The Stochastic Model of Unemployment and Interest Rates

This section sets forth and describes tests of a naive but powerful formulation of the hypothesis that there is a natural rate of unemployment. The hypothesis formulated here is much stricter than the usual statement of the natural-rate hypothesis, which posits that the government can persistently depress the unemployment rate below the "natural rate" only at the cost of accepting an accelerating inflation In contrast, the present formulation implies that there is *no* way that the government can operate so that it can expect to depress the unemployment rate below the natural rate, even in the short run Among other things, that implies that policymakers face no "cruel choice" between inflation and unemployment over *any* relevant time frame

The tests of the natural-rate hypothesis implemented here differ substantially from the usual one, which involves testing the hypothesis that certain sums of distributed-lag weights are unity or zero This usual test has been harshly criticized on theoretical grounds[5] and furthermore is subject to the purely econometric objection that economic time-series data usually yield very little information about "long-run" magnitudes such as the sum of distributed-lag weights [6] The tests implemented here do not seem to depend on estimating any such long-run properties of lag distributions

The present statement of the natural-rate hypothesis is compatible with, but somewhat stronger than, the one presented and tested by Lucas The strategy that I use to test the hypothesis is more naive and purely "statistical" than was Lucas's (1973) procedure, which involved actually estimating a concrete structural model

The Natural-Rate Hypothesis

I begin with the univariate Wold representation of the unemployment rate, Un_t Wold showed that if a variable, for example, Un_t, is an indeterministic, covariance-stationary process, it can be represented as a one-sided moving average of "white noise"

$$Un_t = \sum_{j=0}^{\infty} a_j u_{t-j}, \qquad \sum_{j=0}^{\infty} a_j^2 < \infty, \tag{2 1}$$

where the u's are serially uncorrelated with mean zero and finite variance σ^2 The model in equation (2 1) is obviously intended to apply to deviations of unemployment from its mean and any deterministic components To make things simpler without really altering the essentials, I shall assume that the u's and the other white noises to be introduced below are serially independent [7] I also assume that the roots of

$$\sum_{j=0}^{\infty} a_j \lambda^j = 0$$

---

[5] See Sargent 1971 and Lucas 1972*a*

[6] See Sims 1972*a* A lag distribution that embodies a wrong prior restraint on the sum of the lag weights but is sufficiently flexible can usually achieve a fit arbitrarily close to what could be achieved if the erroneous constraint on the sum of the lag weights were removed. (This assumes that the spectral density of the independent variable has no spike at zero frequency)

[7] Dropping the assumption that the u's and other white noises are serially independent but only serially uncorrelated would necessitate replacing conditional mathematical expectations with linear least-squares forecasts in the subsequent argument With that replacement the argument would go through The statistical tests reported in the next section only utilize the assumption that the various white noises are serially uncorrelated.

lie outside the unit circle, so that Un_t possesses the autoregressive representation

$$Un_t = \sum_{j=1}^{\infty} g_i Un_{t-j} + u_t. \qquad (2.2)$$

Even with these restrictions, equation (2.1) is a very general representation of a covariance-stationary, indeterministic process, the a_j's being chosen to enable the covariogram of $\sum a_j u_{t-j}$ to match that of Un_j. So far, then, I have not restricted the process for the unemployment rate very much.

Let the vector θ_t be the set of observations on all variables observed as of time t or earlier, θ_t includes observations on current and past GNP, interest rates, prices, and any other things, including unemployment itself, thought potentially to contribute to predicting unemployment. The following statement of the natural-rate hypothesis can now be advanced: the unemployment rate Un_t is said to obey the natural-rate hypothesis if in its (univariate) Wold representation (eq. [2.1]), the innovation u_t obeys

$$E(u_t|\theta_{t-1}) = 0, \qquad (2.3)$$

so that the innovation in the unemployment rate is statistically independent of each component of θ_{t-1} and so cannot be predicted on the basis of the information in θ_{t-1}. This means that taking into account components of θ_{t-1} other than lagged Un_t's does not, on the least-squares criterion, improve the forecast of Un_t that can be made on the basis of lagged Un's alone. The least-squares forecast of Un_t on the basis of $Un_{t-1}, Un_{t-2}, \ldots$, call it $\hat{U}n_t$, is given by

$$\hat{U}n_t = \sum_{i=1}^{\infty} g_i Un_{t-i} = \sum_{j=1}^{\infty} a_j u_{t-j}.$$

On our assumption that the u's are serially independent, $\hat{U}n_t = E(Un_t|Un_{t-1}, Un_{t-2}, \ldots) = E(Un_t|u_{t-1}, u_{t-2}, \ldots)$.

The statement that the best forecast of u_t conditional on all past data is simply its unconditional mean of zero amounts to a very strict version of the natural-rate hypothesis. For θ_{t-1} includes past values of monetary and fiscal policy variables. Such variables are asserted to offer no aid in predicting the unemployment rate, once lagged unemployment rates are taken into account. Furthermore, equation (2.3) implies that the current value of any control variable determined via a deterministic feedback rule on θ_{t-1} is also of no use in predicting the unemployment rate. For example, suppose that the logarithm of the money supply at t, m_t, is determined according to the deterministic, very general feedback rule

$$m_t = f(\theta_{t-1}), \qquad (2.4)$$

where f is some (perhaps very complicated) function that determines the monetary authority's feedback rule. Then the above version of the natural-rate hypothesis implies that once lagged Un's are taken into account, current m_t is of no use in predicting Un_t, so that $E(Un_t \mid m_t, Un_{t-1}, Un_{t-2}, \ldots) = E(Un_t \mid Un_{t-1}, Un_{t-2}, \ldots)$. This holds regardless of the nature of the function f or the particular parameter values characterizing f Now feedback rules of the form of equation (2.4) form the class of rules for government-policy variables that control theory indicates to be optimal ones for macroeconometric models (fixed-coefficient stochastic-difference equations) The statement above of the natural-rate hypothesis implies that the choice of f has no effect on the mean of the unemployment rate, conditional on past data. This is a very strong implication about the conditional mean of the unemployment rate, one that denies, for example, that policymakers have any scope to trade a lower expected unemployment rate for a higher expected rate of inflation[8] By way of contrast, the existing macroeconometric models, as usually manipulated, all imply that the parameters of f and the feedback rules for other government-policy variables *do* help determine the conditional mean of the unemployment rate, and that policymakers must face up to a hard choice between the unemployment rate and the inflation rate they can expect to achieve

It is important to note that the definition above of the natural-rate hypothesis does not rule out the possibility that there are correlations between the unemployment rate and other variables such as prices or wages or the money supply It does imply, however, that any such correlations that exist cannot be exploited in predicting the unemployment rate To take an example, Lucas's model of the Phillips curve is

$$Un_t = \sum_{i=1}^{\infty} g_i Un_{t-i} + g_0(X_t - EX_t|\theta_{t-1}) + u'_t,$$

where X_t is the price level at time t, $EX_t|\theta_{t-1}$ is the mathematical expectation of the price level at t, conditional on information available at time $t - 1$, and u'_t is a well-behaved disturbance term, one that satisfies $Eu'_t|\theta_{t-1} = 0$[9] The equation above posits a correlation between the innovations of Un_t and X_t, but notice that

$$E(Un_t|\theta_{t-1}) = \sum_{i=1}^{\infty} g_i Un_{t-i},$$

---

[8] It does not necessarily follow that the distribution of the innovation in unemployment is independent of the feedback rule for policy—only that its conditional mean is The empirical tests reported in this paper are of neutrality-in-conditional-means propositions Stronger neutrality propositions, asserting invariance of the entire probability distribution of some real economic variables with respect to the policy rule, obtain in some macroeconomic models (see, e.g., Sargent and Wallace 1975).

[9] The notations $E_{t-1}X_t$ and $EX_t \mid \theta_{t-1}$ are alternatives denoting the same concept, so that $E_{t-1} X_t \equiv EX_t \mid \theta_{t-1}$.

so that such a correlation does not help in predicting the unemployment rate. Obviously, the same sort of result would obtain were X_t interpreted as a vector of exogenous and endogenous variables.

As another example of correlation between unemployment and another variable that does not aid in forecasting unemployment, consider the system

$$Un_t = \sum_{i=1}^{m} g_i Un_{t-i} + u_t$$

$$X_t = \sum_{i=1}^{n} \lambda_i X_{t-i} + \sum_{i=1}^{q} \gamma_i Un_{t-i} + \varepsilon_t,$$

where the g's, λ's, and γ_i's are parameters, and u_t and ε_t are mutually uncorrelated and serially independent random variables with finite variances In this system, unemployment helps predict X, even taking lagged X's into account, but once lagged Un's are taken into account, lagged values of X are of no aid in predicting unemployment

Testing the Hypothesis

Granger (1969) and Sims (1972*b*) have described the statistical theory that can be used to construct tests of the natural-rate hypothesis as formulated above According to Granger, " We say that Y_t is *causing* X_t if we are better able to predict X_t using all available [past] information than if the information apart from [past] Y_t had been used" (p 428) The formulation above of the natural-rate hypothesis thus posits that the unemployment rate is *caused*, in Granger's sense, by *no* other variables From Granger's paper, a direct statistical test of that hypothesis is available Consider the unemployment rate Un_t and some other variable Y_t Using the method of least squares, estimate the linear regression of Un_t on lagged Un's and lagged Y's as

$$\hat{Un}_t = \sum_{j=1}^{m} \hat{a}_j Un_{t-j} + \sum_{j=1}^{n} \hat{\beta}_j Y_{t-j}, \tag{2 5}$$

where the $\hat{a}_j$'s and $\hat{\beta}_j$'s are least-squares estimates On the null hypothesis that Y does *not* cause Un, the parent parameters β_j, $j = 1$, , n, equal zero The natural-rate hypothesis can then be tested by testing the null hypothesis $\beta_j = 0$ (when $j = 1$, , n), for various choices of Y. Alternatively, lagged values of several variables can be added to the right side of equation (2 5) Or the natural-rate hypothesis, all such variables bear zero coefficients

An alternative way of testing the natural-rate hypothesis as posed here is to employ the test for Granger causality proposed by Sims (1972*b*). Assume that Un and some other series Y are jointly covariance stationary and that they are purely indeterministic. Then the generalization of Wold's representation theorem to n dimensions implies that Un_t and Y_t have the moving-average representation

$$Un_t = \sum_{i=0}^{\infty} a_i \varepsilon_{t-i} + \sum_{i=0}^{\infty} b_i \eta_{t-i} \tag{2.6a}$$

$$Y_t = \sum_{i=0}^{\infty} c_i \varepsilon_{t-i} + \sum_{i=0}^{\infty} d_i \eta_{t-i}, \tag{2.6b}$$

where ε and η are serially uncorrelated and mutually uncorrelated with finite variances, equations (2.6a) and (2.6b) are very general representations of the two processes Un_t and Y_t, the a's, b's, c's, and d's being chosen to make the cross-covariogram between the moving sums on the left-hand sides of the two equations match that between Un and Y. Sims showed that Y does not cause Un in Granger's sense if and only if either all of the a_i's or all of the b_i's in those equations are zero.[10] On the basis of this result, Sims showed that Y_t could be expressed as a one-sided distributed lag of Un_t with a disturbance uncorrelated with past, future, and current Un's if and only if Y fails to "cause" Un. Sims's test for exogeneity of Un is to regress Y on past, present, and future Un's and then to test the null hypothesis that coefficients on future Un's are zero. That is, by least-squares estimate,

$$Y_t = \sum_{i=-n}^{n} \gamma_i Un_{t-i} + e_t,$$

where e_t is a residual. On the null hypothesis that Y does not cause Un, $\gamma_i = 0$ for $i < 0$.

The Interest Rate

The equation for the long-term interest rate is motivated by the rational-expectations version of the expectations theory of the term structure.[11] Let R_{nt} be the yield to maturity on an n-period bond at time t, where n is large in relation to unit increments in t. I approximate the rational-expectations theory of the term structure as asserting

$$R_{nt} = \frac{1}{n}\left(R_{1t} + E_t R_{1t+1} + \cdots + E_t R_{1t+n-1}\right), \tag{2.7}$$

---

[10] This is a very important result, since it establishes the coincidence between Granger causality and the econometrician's definition of statistical exogeneity. (It is assumed that the process $[Un_t/Y_t]$ possesses an autoregressive representation.)

[11] For expositions of the rational-expectations theory of the term structure and evidence that it performs acceptably well, see Shiller (1972) and Modigliani and Shiller (1973).

so that the n-period rate is an average of the current short rate R_{1t} and expected future short rates $E_t R_{1t+j}$, $j = 1, \ldots, n-1$. Expectations about future short rates are assumed to be rational. Subtracting R_{nt} from R_{nt+1} gives $R_{nt+1} - R_{nt} = \eta_{nt+1} + (1/n)(E_t R_{1t+n} - R_{1t})$, where $\eta_{nt+1} = (1/n)[(R_{1t+1} - E_t R_{1t+1}) + (E_{t+1} R_{1t+2} - E_t R_{1t+2}) + \cdots + (E_{t+1} R_{1t+n-1} - E_t R_{1t+n-1})]$. The term η_{nt+1} is of the nature of an "innovation" and as an implication of rationality obeys $E_t \eta_{nt+1} = 0$. Furthermore, for large n and well-behaved (i.e., flat enough) yield curves, $(1/n)(E_t R_{1t+n} - R_{1t}) \simeq 0$. Consequently, for large n, there obtains the approximation

$$E_t R_{nt+1} = R_{nt}, \tag{2.8}$$

which says that the n-period rate is a martingale process.

Suppose that the reduced form for the short-term interest rate is $R_{1t} = \beta Z_t$, where β is conformable to Z_t and where Z_t is a vector of exogenous variables including government expenditures, tax rates, the money supply, and other determinants of the real rate of interest and the expected rate of inflation. That gives us $E_t R_{1t+j} = \beta E_t Z_{t+j}$. Then equation (2.7) becomes $R_{nt} = (1/n)\beta(Z_t + E_t Z_{t+1} + \cdots + E_t Z_{t+n-1})$. So we have

$$\begin{aligned} R_{nt+1} - R_{nt} = (1/n)\beta[(Z_{t+1} - E_t Z_{t+1}) + (E_{t+1} Z_{t+2} - E_t Z_{t+2}) + \cdots \\ + (E_{t+1} Z_{t+n-1} - E_t Z_{t+n-1})] + (1/n)\beta(E_t Z_{t+n} - Z_t) \end{aligned} \tag{2.9}$$

Supposing that Z_t is a vector autoregressive process, it is easy to show that[12]

$$(E_{t+1} Z_{t+j} - E_t Z_{t+j}) = \Gamma_{j-1}(Z_{t+1} - E_t Z_{t+1}), \tag{2.10}$$

where Γ_{j-1} is a square matrix conformable with Z, one whose elements are functions of the parameters of the autoregression for Z. Substituting equation (2.10) into (2.9), we obtain $R_{nt+1} - R_{nt} = (1/n)\beta(I + \Gamma_1 + \Gamma_2 + \cdots + \Gamma_{n-2})(Z_{t+1} - E_t Z_{t+1}) + (1/n)\beta(E_t Z_{t+n} - Z_t)$. Upon imposing our flat-yield-curve approximation $(1/n)(E_t Z_{t+n} - Z_t) = 0$, the equation above becomes

$$R_{nt+1} - R_{nt} = \xi(Z_{t+1} - E_t Z_{t+1}), \tag{2.11}$$

where $\xi = (1/n)\beta(I + \Gamma_1 + \cdots + \Gamma_{n-2})$. This is a version of equation (1.5). As before, we have the implication of equation (2.8), $E_t R_{nt+1} = R_{nt}$.

According to equation (2.8), a regression of $R_{nt+1} - R_{nt}$ against any variables dated t or earlier ought to have coefficients of zero. For example, a regression of $R_{nt+1} - R_{nt}$ against prices or rates of inflation dated t or earlier ought to have zero regression coefficients. The reason is that R_{nt} already has built into it expectations of inflation over almost all of the horizon for R_{nt+1} and that any revisions in those expectations between t and $t + 1$ cannot be predicted on the basis of information available at time t, by virtue of the rationality of those expectations.

---

[12] This is an implication of Wold's chain rule of forecasting (see, e.g., Shiller 1972 and Modigliani and Shiller 1973).

Another way to test equation (2.7) is to note that it implies that R_{nt} is not caused by any variable. That can be tested by fitting two-sided distributed lags of causal candidates against R_{nt} and testing the null hypothesis that the coefficients on future R_n's are zero.

For my purposes, the important implication of the theory is that R_n cannot be predicted better by taking into account other variables, once lagged values of R_n have been taken into account. So it would be perfectly acceptable to modify equation (2.8) to read

$$E_tR_{nt+1} = \sum_{i=0}^{n} w_iR_{nt-i}, \tag{2.8'}$$

which carries the crucial implication that R_{nt} is caused by no other variables. Equation (2.8′) should perhaps be preferred over equation (2.8) according to certain theories about the liquidity premiums that allegedly infest the term structure.[13]

The assertion that other variables such as monetary aggregates and fiscal-policy variables contain no information (over and above that contained in lagged values of the long rate) that can be used to predict the long rate is one that contradicts the implications of all existing macroeconometric models, as they are usually manipulated.[14] Stochastic simulations of these models will in general generate data for which a variety of monetary, fiscal, and other variables "cause" the long rate and thereby aid in its prediction.

---

[13] A more adequate approximation than eq. (2.8) is available, one that does not ignore the term $1/n(E_tR_{1,t+n} - R_{1t})$. Notice that the term-structure eq. (2.7) implies that

$$E_tnR_{nt+1} = (n + 1)R_{n+1,t} - R_{1t} \tag{2.8''}$$

Equation (2.8″) implies that R_n is caused by (i.e., not exogenous with respect to) R_{n+1} and R_1 but is *not* caused by (i.e., is exogenous with respect to) any other variables once R_{n+1} and R_1 are taken into account. Equation (2.8″) shares the classical character of the less adequate approximation equation (2.8). Essentially, (2.8″) asserts that as a block the term structure of interest rates is statistically exogenous or not caused by other variables. This is enough to preserve the classical nature of the model but is weaker than requiring the interest rate on bonds of a given maturity to be statistically exogenous with respect to all other variables.

[14] The St. Louis model is no exception.

Observations on the Tests

The restrictions imposed by the statistical models for unemployment and the interest rate outlined here are stricter than what is really necessary to deliver the classical policy implications of the model. Thus, suppose that X_t is a vector of "real" economic aggregates at time t including variables such as real GNP, unemployment, layoffs, interest rates, and so on, X_t excludes variables measuring the composition of output, such as aggregate consumption and investment and outputs of particular commodities. Let g_t be a list of monetary and fiscal-policy variables at time t. Then a model in general will have classical policy implications if it satisfies

$$E(X_t|X_{t-1}, X_{t-2}, \ldots, g_{t-1}, g_{t-2}, \ldots) = E(X_t|X_{t-1}, X_{t-2}, \ldots), \quad (2.12)$$

so that as a block the aggregate real variables X are statistically exogenous with respect to (not caused by, in Granger's sense) the variables in g. For a system satisfying equation (2.12), movements in the components of g do not have predictable effects on subsequent values of the real variables in X. So equation (2.12) exhibits the same sort of neutrality of certain real variables with respect to monetary and fiscal policy as does the model in Section I.

While the model of Section I is an example of a system satisfying equation (2.12), (2.12) is more general. There are systems satisfying equation (2.12) that violate the hypothesis for the unemployment rate and the interest rate described here in Section II which are key hypotheses underlying the model of Section I. Thus, equation (2.12) does *not* imply $E(Un_t|Un_{t-1}, Un_{t-2}, \ldots, g_{t-1}, g_{t-2}, \ldots) = E(Un_t|Un_{t-1}, Un_{t-2}, \ldots)$, even though Un_t is a component of X_t. A simple example that illustrates this is a system satisfying equation (2.12) in which, say, layoffs help cause unemployment. Suppose that some components of g_t are set via a feedback rule on layoffs. Then even though g does not cause (help predict) Un when lagged unemployment *and* lagged layoffs are taken into account, components of g will help predict unemployment when only lagged unemployment is taken into account. This is because g contains some information about lagged layoffs. This is a "spurious" type of causality from g to Un in which an omitted variable (layoffs) is causing both g and Un (see Granger 1969), when layoffs are omitted, g only appears to cause Un because it is standing in for the omitted lagged-layoff rates.

The possibility of such spurious apparent causality running from components of g to Un is noteworthy, since the statement above of the natural-rate hypothesis is so very strict. In particular, it rules out even the possibility that other real variables (the components of X in eq. [2.12]) cause unemployment. This seems too drastic, since it is easy to imagine structures in which there is extensive causality from, say, GNP and layoffs to unemployment that satisfy equation (2.12) and so are basically classical in nature For such a system, our tests might well reject the very strict version of the natural-rate hypothesis adopted above

While failure of monetary and fiscal-policy variables to cause unemployment and other real variables is sufficient to deliver classical policy implications, it is not really necessary One can imagine structures in which policy variables cause (help predict) unemployment and other real variables, but in which switching from one deterministic rule for setting the policy variable to another leaves the stochastic behavior of unemployment unchanged As an example, consider the structural system

$$Un_t = \sum_{i=1}^{n_1} \lambda_i Un_{t-i} + \beta_0(m_t - E_{t-1}m_t) + \beta_1(m_{t-1} - E_{t-2}m_{t-1}) + u_t \tag{2 13}$$

and

$$m_t = \sum_{i=1}^{n_2} \delta_i m_{t-i} + \varepsilon_t, \tag{2 14}$$

where ε_t and u_t are random variables, and $E_{t-1}\varepsilon_t = E_{t-1}u_t = 0$ For the structure above, it is easy to calculate

$$E(Un_t|Un_{t-1}, \quad , m_{t-1}, m_{t-2}, \quad) = \sum_{i=1}^{n_1} \lambda_i Un_{t-i} + \beta_1\Bigl(m_{t-1} - \sum_{i=1}^{n_2} \delta_i m_{t-i-1}\Bigr)$$

It follows that m helps predict (causes) Un_t But notice that according to equation (2 13) switching from one deterministic rule for m (i e., a rule for which $m_t = E_{t-1}m_t$) to any other deterministic rule will leave the stochastic behavior of unemployment unaltered Even though m causes unemployment in this system, it is true that one deterministic rule is as good as any other, so that there is no scope for countercyclical policy by way of "leaning against the wind "

The preceding observations suggest reasons for believing that this paper tests versions of classical hypotheses that are really stronger than what is necessary to deliver classical policy conclusions, so that the tests seem biased against the natural-rate hypothesis and other classical hypotheses However, it is important to note that the tests are not uniformly biased against classical hypotheses, since it is possible to concoct nonclassical systems that will mimic the classical characteristics that my tests look for. Thus, the tests might be fooled into failing to reject the natural-rate hypothesis in a system for which that hypothesis is false Suppose that the true reduced form for Un_t is

$$Un_t = \sum_{i=1}^{2} \lambda_i Un_{t-i} + \alpha_0 m_t + \varepsilon_t, \tag{2.15}$$

where $E(\varepsilon_t | Un_{t-1}, \quad , m_t, m_{t-1}) = 0$ and where the λ's and α's are fixed parameters Suppose that the authority sets m_t according to the deterministic feedback rule

$$m_t = \sum_{i=1}^{2} \delta_i Un_{t-i}$$

Then clearly

$$E(Un_t | Un_{t-1}, Un_{t-2}, \quad , m_t, m_{t-1}) = \sum_{i=1}^{2} (\lambda_i + \alpha_0 \delta_i)\, Un_{t-i},$$

Here Un is not caused by m, in Granger's sense, because the authority, by making m_t an exact function of past Un's, eliminates any value from the m series for predicting Un

While the tests might be fooled by such a structure, that structure itself seems unlikely to me In particular, if the reduced form were equation (2 15) and the authority were to set m_t by a feedback only on lagged Un's, and not also on other variables, presumably the authority would want to minimize the variance of Un_t, which it could accomplish by eliminating any serial correlation in Un_t That is, in our example, it could minimize the variance in Un by setting $\lambda_1 + \alpha_0\delta_1 = 0, \lambda_2 + \alpha_0\delta_2 = 0$ Then the variance of unemployment would equal the variance of ε_t But in reality, variables like unemployment and the deviation of GNP from trend are highly serially correlated That makes it hard to believe that any failure of, say, m to cause Un is due to the authority's manipulating m in response to past movements in Un, since that requires imputing to the authority a perverse objective, that is, one tolerating much serial correlation and variance in Un

III. Empirical Results

Tables 1-6 report the results of performing tests along the lines proposed by Granger and Sims for quarterly data on the dependent variables spanning the period 1952 II-1972 III The unemployment rate for all civilian workers is used for Un, while Moody's Baa corporate bond index is taken for the long-term interest rate R The variables used as candidates for the "causal" variables Y are the logarithm of the money supply, currency plus demand deposits, (m), the federal, state, and local government surplus on the national-income-accounts basis in 1958 dollars

($surp$), the logarithms of the GNP deflator (p); a straight-time wage index in manufacturing (w), federal, state, and local purchases of goods and services in 1958 dollars (g); and federal, state, and local purchases in current dollars ($g\$$).[15]

Each of the series has been seasonally adjusted by taking the Fourier transform of the series, setting its real and imaginary parts to zero in a band of width $\pi/12$ about the seasonal frequencies, and then taking the inverse Fourier transform to obtain a seasonally adjusted series.[16] This method has the virtue of applying a seasonal-adjustment filter with the same frequency-response function to each series, thereby avoiding the distortions in estimating distributed lags between variables that can be caused where the series have been adjusted asymmetrically (see Sims 1974 and Wallis 1974). Furthermore, the method reduces the spectral power of the series to zero at the seasonal frequencies, which Sims (1974) has argued helps eliminate bias in the form of seasonal patterns showing up in estimated distributed-lag coefficients.

Table 1 reports the results of implementing Granger's test for causality between Un and each of the Y candidates listed above. For each Y, the test is run in both directions: first Un is regressed on lagged Un's and lagged Y's to permit testing the null hypothesis that Y does not cause Un (i.e., that the coefficients on lagged Y's are zero), then Y is regressed on lagged Y's and lagged Un's to permit testing the null hypothesis that Un does not cause Y (i.e., that the coefficients on lagged Un's are zero). Regressions in both directions include a constant and a linear trend. The regressions include four lagged values of the dependent variable and six lagged values of the other variable. The F-statistic pertinent for

---

[15] The wage is an index of the straight-time manufacturing wage (w), which is seasonally adjusted and reported on a monthly basis in *Employment and Earnings* (Department of Labor, Bureau of Labor Statistics). The civilian unemployment rate (Un) seasonally unadjusted, on a monthly basis, was taken from *Employment and Earnings*. For population I used the civilian noninstitutional population aged 16 and over, constructed by subtracting armed forces numbers from the total population aged 16 and over. The noninstitutional population aged 16 and over was interpolated from annual figures compiled by the Current Population Survey and reported in table 1, Bureau of Labor Statistics *Handbook of Labor Statistics*, 1973. Armed forces numbers were obtained by averaging monthly numbers reported in *Employment and Earnings*. The civilian labor force aged 16 years and older was taken from *Employment and Earnings* and divided by pop, to obtain the labor force participation rate. The money supply (m) is $M1$, currency plus adjusted demand deposits taken from the *Federal Reserve Bulletin*. The Baa rate (R) was obtained from Moody's Investor's Service. For R, w, and Un, the monthly figures were averaged to obtain quarterly figures. The GNP deflator (p), federal, state, and local purchases in current dollars ($g\$$), and federal, state, and local purchases in 1958 dollars (g) were all taken from the National Income Accounts, the federal, state, and local government surplus in current dollars was also taken from the National Income Accounts and then divided by the GNP deflator to obtain the surplus in 1958 dollars ($surp$).

[16] A deterministic trend was extracted before taking the Fourier transform and then added back in after taking the inverse transform. The degrees of freedom for the F-statistics have been adjusted for the loss of degrees of freedom due to setting the seasonal bands to zero. The appropriate correction is described by Sims (1974).

TABLE 1

REGRESSION OF $Y(t) = \sum_{s=1}^{4} \alpha(s)Y(t-s) + \sum_{l=1}^{6} \beta(l)X(t-l) + \delta_1 t + \delta_0$

1952II–1972III

Y	X	$\alpha(1)$	$\alpha(2)$	$\alpha(3)$	$\alpha(4)$	$\beta(1)$	$\beta(2)$	$\beta(3)$	$\beta(4)$	$\beta(5)$	$\beta(6)$	δ_1	δ_0	$\bar{R}^2$	SE of Est Adj	D-W	F-Statistic on All β $F(6, 60)$
U_n	m	1 317 (10 25)	−0 541 (−2 64)	−0 029 (−0 14)	0 113 (0 94)	−14 244 (−1 89)	7 787 (0 57)	7 578 (0 53)	−24 943 (−1 75)	26 849 (1 85)	−0 862 (−0 10)	−0 010 (−1 32)	−9 526 (−1 86)	918	0 346	1 973	2.627*
m	U_n	1 279 (9 75)	−0 367 (−1 60)	0 022 (0 09)	0 063 (0 44)	−0 002 (−0 88)	0 003 (0 82)	0 0005 (0 13)	0 00002 (0 01)	−0 002 (−0 56)	−0 0009 (−0 44)	0 0002 (1 33)	0 022 (0 26)	999	0 006	1 991	1 506
U_n	swp	1 435 (10 84)	−0 540 (−2 33)	−0 121 (−0 52)	0 134 (1 02)	−0 020 (−1 32)	0 035 (1 58)	−0 035 (−1 56)	0 029 (1 33)	−0 002 (−0 07)	−0 007 (−0 57)	−0 0004 (−0 21)	0 439 (2 36)	906	0 372	1 995	0.954
swp	U_n	1 143 (8 60)	−0 399 (−2 02)	0 149 (0 75)	−0 248 (−1 90)	−0 319 (−0 26)	0 268 (0 13)	−0 848 (−0 40)	2 875 (1 36)	−3 330 (−1 73)	1 979 (1 77)	−0 017 (−1 04)	−3 200 (−1 84)	814	3 323	1.992	1 034
U_n	g	1 498 (11 99)	−0 641 (−2 87)	−0 027 (−0 12)	0 057 (0 40)	0 814 (0 75)	−1 085 (−0 82)	−0 037 (−0 03)	1 152 (0 88)	−1 336 (−1 14)	0 143 (0 20)	0 002 (0 28)	2 036 (0 44)	902	0 380	1.917	0 489
g	U_n	0 685 (5 20)	−0 042 (−0 27)	0 069 (0 46)	−0 110 (−0 96)	−0 016 (−1 10)	−0 004 (−0 15)	0 024 (0 88)	−0 065 (−2 42)	0 034 (1 31)	0 004 (0 22)	0 003 (3 28)	1 840 (3 30)	960	0 043	1 979	4.564**
U_n	gS	1 506 (12 28)	−0 614 (−2 76)	−0 134 (−0.57)	0 147 (1 09)	0 204 (0 07)	−4 415 (−0 95)	7 494 (1 73)	0 211 (0 05)	−6 820 (−1 61)	3 455 (1 56)	−0 002 (−0 19)	−0 052 (−0 02)	908	0 368	2 056	1 145
gS	U_n	1 135 (9 84)	−0 066 (−0 37)	0 099 (0 57)	−0 292 (−2 99)	−0 014 (−2 87)	0 021 (2 45)	−0 006 (−0 63)	−0 008 (−0 86)	0 004 (0 43)	−0 003 −(0 55)	0 002 (3 99)	0 530 (3 84)	999	0 014	1 806	2 504*
U_n	p	1 469 (11.55)	−0 610 (−2 73)	−0 120 (−0 54)	0 132 (1 04)	−15 845 (−1 34)	22 334 (1 23)	−0 608 (−0 03)	−12 832 (−0 71)	14 059 (0 85)	−3 717 (−0 34)	−0 018 (−1 48)	−14 386 (−1 43)	906	0 372	2 063	0 936
p	U_n	1 090 (8 42)	0 003 (0 02)	0 034 (0 17)	−0 120 (−0 87)	0 0004 (0 31)	−0 002 (−0 94)	0 002 (0 78)	−0 002 (−0 88)	0 001 (0 54)	−0 0004 (−0.31)	0 00002 (0.20)	−0 026 (−0 30)	999	0 004	2 019	1 222
U_n	w	1 238 (8 02)	−0 437 (−1 94)	−0 204 (−0 92)	0 122 (0 95)	−18 346 (−1 76)	25 905 (1 81)	−20 746 (−1 44)	20 274 (1 46)	1 146 (0 08)	1 050 (0 10)	−0 084 (−2 44)	−2 469 (−2 36)	917	0 350	2 027	2 371*
w	U_n	1 066 (7 05)	−0 347 (−1 66)	0 431 (2 03)	−0 074 (−0 43)	−0 003 (−1 50)	0 00007 (0 02)	0 00001 (0 00)	0 004 (1 08)	−0 006 (−1 88)	0 002 (1 15)	−0 0006 (−1 72)	−0 010 (−0 79)	9996	0 005	1 969	1 638

NOTE.—t-statistics for coefficients appear in parentheses below relevant coefficients.
* Significant at 5%
** Significant at 1%

testing the null hypothesis that the dependent variable is not caused by the other variable is reported in the last column.

The F-statistic for m as the causal variable influencing Un is significant at the 95 percent confidence level, though not at the 99 percent level. Similarly, the F-statistic for w as a causal variable for Un is significant at the 95 percent confidence level. None of the other causal candidates obtains an F that would require rejecting the null hypothesis that they do not cause unemployment. In particular, notice that the GNP deflator does not appear to cause unemployment.

In the other direction, the F-statistics reveal that the hypothesis that Un does not cause g or $g\$$ can be rejected at the 95 percent confidence level. The hypothesis that Un does not cause the other four variables cannot be rejected.

Tables 2 and 3 report summary statistics for the regressions implementing Sims's test for unemployment.[17] Two-sided distributed lags were calculated in each direction, one with Un as the dependent variable and the causal candidate Y as the "independent" variable, the other with Un and Y reversed. The data were quasi-differenced by applying the filter $(1 - .75L)^2$. Each regression included a constant and a trend, with four lead variables and 12 lagged variables. The regressions were first estimated by the method of least squares. Then the Fourier transform of the distributed-lag coefficients was calculated. The amplitude of the Fourier transform was inspected to see if peaks occurred at the seasonal frequencies. In those cases where a peak occurred, indicating a seasonal pattern in the coefficients, the regressions were recomputed using Theil's mixed estimator to incorporate weak, stochastic prior information stating that there is no seasonal pattern in the distributed lag. In particular, suppose the regression estimated is

$$Un_t = \sum_{i=-4}^{12} b_i Y_{t-i} + \text{residual}_t,$$

and that a seasonal pattern characterizes the b_i's. The regression was then recalculated by adding observations on the three constraints

$$\begin{aligned} b_{-4} + b_0 + b_4 + b_8 &= b_{-3} + b_1 + b_5 + b_9 + U_1, \\ b_{-4} + b_0 + b_4 + b_8 &= b_{-2} + b_2 + b_6 + b_{10} + U_2, \\ b_{-4} + b_0 + b_4 + b_8 &= b_{-1} + b_3 + b_7 + b_{11} + U_3, \end{aligned}$$

where the U's are random variables obeying $EU_1 = EU_2 = EU_3 = 0$. Theil's mixed estimator requires estimates of the standard error of the

---

[17] To save space, the graphed lag distributions and various summary statistics of the two-sided tests have been relegated to a mimeographed appendix that is available from the author on request. The graphs and various statistics from the Hannan efficient regressions discussed below are also in this appendix.

TABLE 2

F-STATISTIC—TWO-SIDED TESTS

VARIABLE NAME (Y)	INDEPENDENT VARIABLE Un* (1)	 Y† (2)
m	0 401‡	0 951§
$surp$	1 09‡	0 396§
g	0 344‖	2 854‖
$g\$$	0 374‡	1 255‡
p	0 647‖	0 479§
w	1 472‖	1 238§

NOTE —All F's are $F(4, 50)$, significance levels are 2 56 for 95% confidence, 3 72 for 99% confidence

Col 1 regressions $Y_t = \sum_{i=-4}^{12} w_i Un_{t-i}$, col 2 regressions $Un_t = \sum_{i=-4}^{12} w'_i Y_{t-i}$

* F-statistic is pertinent for testing null hypothesis $w_{-4} = w_{-3} = w_{-2} = w_{-1} = 0$
† F-statistic is pertinent for testing null hypothesis $w'_{-4} = w'_{-3} = w'_{-2} = w'_{-1} = 0$
‡ Theil constraint used with $\sigma_u = \max w_i - \min w_i$
§ No Theil constraint used
‖ Theil constraint used with $\sigma_u = (\max w_i - \min w_i)/2$

disturbances in the regression and the standard errors of U_1, U_2, and U_3 The former was taken as equal to the standard error of the residuals in the original least-squares regression The latter standard errors were taken as equal to one another at σ_u, which was set at either $(\max_i \hat{b}_i - \min_i \hat{b}_i)$ or $(\max_i \hat{b}_i - \min_i \hat{b}_i)/2$, where the $\hat{b}_i$'s are from the original least-squares regression The covariance of each U with all other random variables was assumed to be zero Estimation incorporating this prior information in most cases sufficed to eliminate the seasonal in the distributed-lag coefficients

Table 2 summarizes the F-statistics pertinent for testing the null hypothesis that the coefficients on future values of the right-side variable are zero, that is, the null hypothesis that the left-side variable does not cause the right-side variable For no causal candidate Y does the F-statistic indicate rejecting that Y does not cause Un at the 95 percent confidence level In particular, notice that in contrast to the results from applying the direct Granger test, it is not possible to reject the hypothesis that m or w does not cause Un In the other direction, the F-statistic reveals that the hypothesis that g is not caused by Un must be rejected at the 95 percent confidence level The next highest F is for $g\$$, though it is not significant at the 95 percent confidence level Qualitatively, the overall pattern of the results is similar to that obtained by applying Granger's test, with the important exceptions of the different results rendered for whether m causes Un and for whether w causes Un

Table 3 reports F-statistics pertinent for testing whether the coefficients on current and lagged right-hand side variables are zero in the one-sided regressions corresponding to those in table 2. Only the F-statistics

TABLE 3

F-STATISTICS FOR COEFFICIENTS ON CURRENT AND LAGGED VARIABLES·

I $Y_t = \sum_{i=0}^{12} \alpha_i Un_{t-i} + \beta_0 + B_1 t$, II $Un_t = \sum_{i=0}^{12} \gamma_i Y_{t-i} + \delta_0 + \delta_1 t$

VARIABLE NAME (Y_t)	INDEPENDENT VARIABLE *Un*	 Y_t
m	0 883^a	0 911^b
surp	1 685^a	1 999^{*b}
g	1 454^c	0 520^c
g$	1 785^a	1 441^a
p	0 571^c	1 347^b
w	1 003^c	2 355^{*b}

NOTE —All *F*'s are $F(13, 54)$, all data are filtered $(1 - 75L)^2$

[a] Theil constraint used with $\sigma_u = \max w_i - \min w_j$

[b] No Theil constraint used

[c] Theil constraint used with $\sigma_u = (\max w_i - \min w_i)/2$

[*] Significant at 5%

for the regression of *Un* on *surp* and *Un* on *w* are significant at the 95 percent confidence level

Tables 4, 5, and 6 report the results of applying Granger's and Sims's tests to determine whether the long-term interest rate, as measured by the Baa yield index, is statistically exogenous as implied by our theory Table 4 records the results of applying the direct Granger test The *F*-statistic is the one pertinent for testing that the coefficients on lagged values of the causal candidate *Y* are all zero, so that *Y* does not cause or help predict the dependent variable Where *R*Baa is the dependent variable, *w* is the only causal candidate that obtains an *F*-statistic that is significant at the 95 percent confidence level At that confidence level, the results are thus consistent with the implications of the theory, with the exception of the results for *w*, which indicate that *w* causes *R*Baa In the reverse direction, the hypothesis that *R*Baa does not cause the money supply must be rejected at the 95 percent confidence level

For Sims's test, table 5 summarizes the *F*-statistics pertinent for testing the null hypothesis of no causality for the interest rate The results are compatible with those obtained from applying Granger's test The hypothesis that *R*Baa is not caused by the causal candidate can be rejected at the 95 percent confidence level only for *w* In the reverse direction, the hypothesis that *R*Baa fails to cause *m* must be rejected at the 95 percent confidence level

Table 6 reports the *F*-statistics pertinent for testing the null hypothesis that coefficients on current and lagged values of the causal candidates are zero in the one-sided regressions corresponding to those in table 5 The *F*'s for *w* on *R*Baa and *R*Baa on *w* are the only ones significant at the 95 percent confidence level, though a couple of others are marginal

TABLE 4

REGRESSION OF $Y(t) = \sum_{s=1}^{4} \alpha(s)Y(t-s) + \sum_{l=1}^{6} \beta(l)X(t-l) + \delta_1 t + \delta_0$

1952II-1972III

Y	X	$\alpha(1)$	$\alpha(2)$	$\alpha(3)$	$\alpha(4)$	$\beta(1)$	$\beta(2)$	$\beta(3)$	$\beta(4)$	$\beta(5)$	$\beta(6)$	δ_1	δ_0	$\bar{R}^2$	SE of Est. Adj	D-W	F-Statistic on All β $F(6, 60)$
RBaa	*m*	1.447 (11.54)	−0.734 (−3.26)	0.553 (2.41)	−0.378 (−2.87)	0.635 (0.17)	3.311 (0.50)	−2.633 (−0.39)	−4.584 (−0.68)	5.248 (0.79)	−1.016 (−0.27)	−0.0004 (−0.11)	−4.251 (−1.29)	.992	0.161	1.867	0.805
m	*RBaa*	1.337 (10.39)	−0.187 (−0.80)	−0.517 (−1.27)	0.194 (1.38)	−0.012 (−2.55)	0.005 (0.62)	0.017 (1.89)	−0.011 (−1.19)	−0.002 (−0.21)	0.001 (0.21)	0.00001 (0.14)	−0.129 (−1.30)	.999	0.006	1.925	2.730*
RBaa	*nrp*	1.471 (11.45)	−0.715 (−3.16)	0.491 (2.14)	−0.290 (−2.12)	0.009 (1.48)	−0.011 (−1.15)	0.010 (1.06)	−0.007 (−0.74)	0.002 (0.19)	0.001 (0.19)	0.003 (1.61)	0.141 (1.58)	.992	0.163	1.867	0.593
nrp	*RBaa*	1.107 (8.83)	−0.331 (−1.77)	0.156 (0.85)	−0.265 (−2.18)	0.889 (0.34)	−7.705 (−1.66)	9.335 (1.89)	−7.956 (−1.58)	5.698 (1.17)	−1.244 (−0.44)	0.053 (1.41)	2.155 (1.19)	.828	3.194	1.960	1.948
RBaa	*g*	1.485 (11.82)	−0.720 (−3.20)	0.423 (1.88)	−0.254 (−2.00)	−0.044 (−0.10)	0.110 (0.20)	0.421 (0.75)	−0.426 (−0.76)	−0.030 (−0.06)	0.276 (0.90)	0.002 (0.72)	−1.139 (−0.97)	.992	0.164	1.894	0.462
g	*RBaa*	0.918 (7.15)	−0.080 (−0.46)	0.036 (0.21)	−0.096 (−0.76)	0.060 (1.49)	−0.085 (−1.15)	0.067 (0.86)	−0.070 (−0.88)	0.047 (0.63)	−0.025 (−0.59)	0.002 (2.51)	0.990 (2.92)	.946	0.050	1.971	0.711
RBaa	*g$*	1.485 (11.75)	−0.758 (−3.40)	0.470 (2.11)	−0.282 (−2.25)	1.130 (0.90)	−1.303 (−0.67)	1.216 (0.64)	−1.598 (−0.84)	1.161 (0.62)	−0.145 (−0.15)	−0.002 (−0.50)	−1.640 (−1.27)	.992	0.164	1.888	0.502
g$	*RBaa*	1.193 (10.38)	−0.151 (−0.82)	0.218 (1.22)	−0.352 (−3.61)	−0.003 (−0.21)	0.026 (1.19)	−0.039 (−1.63)	0.024 (0.98)	−0.012 (−0.53)	0.007 (0.52)	0.001 (3.60)	0.369 (3.31)	.999	0.015	1.650	0.656
RBaa	*p*	1.388 (10.86)	−0.687 (−3.20)	0.375 (1.74)	−0.187 (−1.30)	5.516 (1.11)	−3.821 (−0.50)	3.812 (0.50)	6.530 (0.88)	−11.143 (−1.62)	−1.321 (−0.29)	0.007 (0.92)	2.144 (0.24)	.992	0.156	1.962	1.589
p	*RBaa*	1.083 (8.24)	0.003 (0.02)	0.021 (0.11)	−0.106 (−0.78)	0.003 (0.99)	−0.004 (−0.75)	0.010 (1.57)	−0.008 (−1.31)	0.0003 (0.04)	−0.0002 (−0.06)	0.00002 (0.09)	−0.004 (−0.02)	.999	0.004	2.016	1.482
RBaa	*w*	1.531 (11.87)	−0.770 (−3.45)	0.267 (1.21)	−0.117 (−0.92)	7.898 (2.19)	−15.196 (−2.66)	15.141 (2.48)	0.587 (0.09)	−12.076 (−2.00)	3.635 (0.93)	0.005 (0.60)	0.154 (0.36)	.993	0.146	2.017	3.199**
w	*RBaa*	1.148 (8.95)	−0.333 (−1.65)	0.361 (1.70)	−0.282 (−2.04)	0.004 (0.94)	0.002 (0.24)	−0.004 (−0.45)	−0.004 (−0.52)	0.005 (0.63)	0.002 (0.45)	0.001 (2.20)	0.038 (2.45)	.9996	0.005	2.065	1.704

NOTE.—t-statistics for coefficients appear in parentheses below relevant coefficients.
* Significant at 5%.
** Significant at 1%.

TABLE 5

F-STATISTICS—TWO-SIDED TESTS

VARIABLE NAME (Y)	INDEPENDENT VARIABLE RBaa* (1)	 $Y^{\dagger}$ (2)
m	0 886‡	2 808‡
$surp$	0 708‡	1 454‡
g	0 285§	0 373‡
$g\$$	0 853§	0 661‡
p	1 339§	0 450§
w	3 251§	1 932§

NOTE – All F's are $F(4, 50)$, significance levels are 2 56 for 95% confidence, 3 72 for 99% confidence

Col 1 regressions $Y_t = \sum_{i=-4}^{12} w_i RBaa_{t-i}$, col 2 regressions $RBaa_t = \sum_{i=-4}^{12} w_i Y_{t-i}$

* F-statistic is pertinent for testing null hypothesis $w_{-4} = w_{-3} = w_{-2} = w_{-1} = 0$
† F-statistic is pertinent for testing null hypothesis $w_{-4} = w'_{-3} = w_{-2} = w'_{-1} = 0$
‡ No Theil constraint used
§ Theil constraint used with $\sigma_u = (\max w_i - \min w_i)/2$

TABLE 6

F-STATISTICS FOR COEFFICIENTS ON CURRENT AND LAGGED VARIABLES

$$\text{I}\quad Y_t = \sum_{i=0}^{12} \alpha_i RBaa_{t-i} + \beta_0 + \beta_1 t \text{ (SEASONAL DUMMIES INCLUDED),}$$

$$\text{II}\quad RBaa_t = \sum_{i=0}^{12} \gamma_i Y_{t-i} + \delta_0 + \delta_1 t$$

VARIABLE NAME (Y_t)	INDEPENDENT VARIABLE RBaa (1)	 Y_t (2)
m	1 854^{a}	0 511^{a}
$surp$	1 771^{a}	1 648^{a}
g	0 344^{b}	0 446^{a}
$g\$$	0 798^{b}	0 408^{a}
p	1 751^{b}	1 025^{b}
w	2 554^{**b}	2 004^{*b}

NOTE – All F's are $F(13, 54)$, all data are filtered with $(1 - 75L)^2$
^{a} No Theil constraint used
^{b} Theil constraint used with $\sigma_u = (\max w_i - \min w_i)/2$
* Significant at 5%
** Significant at 1%

and may be understated because possibly too many lagged variables have been included

Table 7 reports F-statistics pertinent for testing whether the labor force participation rate nf is exogenous with respect to various causal candidates The model implies that n_{f_t} is exogenous with respect to all variables in the model, with the possible exception of the unemployment rate. The unemployment rate can cause the labor force participation rate, say through the "discouraged-worker effect," while not destroying the

TABLE 7

F-STATISTICS—TWO-SIDED TESTS

VARIABLE NAME (Y)	INDEPENDENT VARIABLE nf* (1)	 Y† (2)
Un	3.060‡	0.945§
m	1.591§	0.390‡
surp	1.320‡	0.514§
g	1.471§‖	1.594‡
g$	0.499‡	0.819§‖
p	0.901‡	1.487‡
w	0.586‡	1.498§‖

NOTE.—All F's are $F(4, 50)$, significance levels are 2.56 for 95% confidence, 3.72 for 99% confidence.

Col 1 regressions $Y_t = \sum_{i=-4}^{12} w_i nf_{t-i}$, col 2 regressions $nf_t = \sum_{i=-4}^{12} w'_i Y_{t-i}$

* F-statistic is pertinent for testing null hypothesis $w_{-4} = w_{-3} = w_{-2} = w_{-1} = 0$

† F-statistic is pertinent for testing null hypothesis $w'_{-4} = w'_{-3} = w'_{-2} = w'_{-1} = 0$

‡ No Theil constraint used

§ Theil constraint used with $\sigma_u = (\max w_i - \min w_i)/2$

‖ Some seasonal remains in the distributed lag weights despite the imposition of Theil smoothness prior

"recursive" structure of the model which prevents monetary and fiscal policy variables from causing the real variables *Un*, *nf*, and *y*.

The *F*-statistics in table 7 emerge from implementing Sims's test. The only *F*-statistic that is significant at the 95 percent confidence level is the one pertinent for testing the null hypothesis that *Un* fails to cause *nf*. At that significance level the null hypothesis must be rejected, which is compatible with the presence of a discouraged-worker effect that is useful for predicting labor force participation. While none of the other *F*-statistics is significant, the regression of m_t against nf_t did obtain several large and statistically significant coefficients on leading values of *nf*. This indicates that one ought perhaps to be cautious about the null hypothesis that *m* does not cause *nf*, despite the insignificant *F*-statistic. With this possible exception, the regressions summarized in table 7 are consistent with the causal structure imposed by the model upon *nf*.

Table 8 reports the results of applying Granger's test to *nf* and various causal candidates. At the 95 percent confidence level, *nf* appears to cause *w*, *p*, *g*, and *Un*, while only *Un* appears to cause *nf*.[18]

---

[18] While the Durbin-Watson statistics from most of the two-sided regressions are close to two, there is a possibility that the presence of higher than first-order serial correlation is making inappropriate the *F*-statistics in the text. For this reason, the two-sided regressions were recomputed using a version of Hannan's efficient estimator, which is asymptotically equivalent to generalized least squares allowing for high-order serial correlation in the disturbances. The results are reported in the mimeographed appendix to this paper (n. 17 above). The general pattern agrees with the results in the text, though there are differences in details. For example, in the Hannan efficient results, *w* does not seem to cause the Baa rate or the unemployment rate. If anything, then, the Hannan efficient regressions seem more favorable to the exogeneity hypotheses imposed by the classical model than are the two-sided regressions reported in the text.

TABLE 8

REGRESSION OF $Y(t) = \sum_{s=1}^{8} \alpha(s)Y(t-s) + \sum_{l=1}^{6} \beta(l)X(t-l) + \delta_1 t + \delta_0$

1952II–1972III

Y	X	$\alpha(1)$	$\alpha(2)$	$\alpha(3)$	$\alpha(4)$	$\alpha(5)$	$\alpha(6)$	$\alpha(7)$	$\alpha(8)$	$\beta(1)$	$\beta(2)$
n_f	U_n	1.027	−0.151	−0.063	0.279	−0.227	0.101	−0.147	0.115	−0.0007	−0.0005
		(8.14)	(−0.83)	(−0.34)	(1.50)	(−1.24)	(0.60)	(−0.89)	(0.95)	(−0.52)	(−0.20)
U_n	n_f	1.540	−0.757	0.159	−0.390	0.430	0.017	−0.162	0.069	−21.500	22.961
		(12.02)	(−3.22)	(0.64)	(−1.57)	(1.68)	(0.07)	(−0.75)	(0.57)	(−1.75)	(1.29)
n_f	m	0.919	−0.176	0.090	−0.086	0.190	−0.116	−0.101	0.028	−0.057	0.002
		(6.76)	(−0.99)	(0.50)	(−0.49)	(1.08)	(−0.67)	(−0.60)	(0.22)	(−0.64)	(0.01)
m	n_f	1.441	−0.525	−0.117	0.465	−0.474	0.352	−0.228	0.121	−0.036	−0.073
		(10.29)	(−2.01)	(−0.41)	(1.62)	(−1.60)	(1.10)	(−0.71)	(0.70)	(−0.17)	(−0.25)
n_f	rwp	0.927	−0.086	0.028	0.114	0.022	0.013	−0.131	0.027	0.00009	0.00003
		(6.92)	(−0.48)	(0.16)	(0.64)	(0.12)	(0.08)	(−0.74)	(0.20)	(0.55)	(0.11)
rwp	n_f	1.089	−0.374	0.163	−0.398	0.326	−0.210	−0.102	−0.024	−22.533	−144.945
		(8.34)	(−1.93)	(0.81)	(−2.02)	(1.74)	(−1.13)	(−0.58)	(−0.20)	(−0.22)	(−1.06)
n_f	g	0.943	−0.071	−0.031	0.143	0.081	−0.119	−0.155	0.105	0.013	−0.026
		(7.42)	(−0.41)	(−0.18)	(0.84)	(0.46)	(−0.64)	(−0.85)	(0.78)	(1.19)	(−1.72)
g	n_f	0.931	−0.087	0.052	−0.129	0.140	−0.090	0.081	−0.147	1.917	1.033
		(7.19)	(−0.51)	(0.31)	(−0.78)	(0.92)	(−0.69)	(0.62)	(−1.57)	(1.25)	(0.50)
n_f	$g\$$	0.969	−0.128	0.014	0.037	0.022	0.046	−0.144	0.052	0.046	−0.047
		(7.27)	(−0.73)	(0.08)	(0.21)	(0.13)	(0.27)	(−0.84)	(0.41)	(1.45)	(−0.94)
$g\$$	n_f	1.082	−0.100	0.242	−0.438	0.182	−0.052	0.052	−0.098	0.440	0.553
		(8.45)	(−0.52)	(1.26)	(−2.35)	(0.93)	(−0.27)	(0.29)	(−1.04)	(0.89)	(0.80)
n_f	p	0.938	−0.185	−0.011	0.039	0.050	−0.085	−0.179	0.098	0.181	−0.078
		(6.73)	(−1.02)	(−0.06)	(0.22)	(0.28)	(−0.47)	(−0.99)	(0.66)	(1.18)	(−0.37)
p	n_f	0.853	0.038	0.138	0.196	−0.185	0.086	−0.145	−0.048	0.360	−0.107
		(6.61)	(0.22)	(0.79)	(1.15)	(−1.14)	(0.51)	(−0.84)	(−0.45)	(3.10)	(−0.68)
n_f	w	0.920	−0.175	−0.020	0.021	0.075	−0.091	−0.173	0.081	0.207	−0.073
		(6.93)	(−0.95)	(−0.11)	(0.11)	(0.40)	(−0.50)	(−0.96)	(0.57)	(1.85)	(−0.42)
w	n_f	1.049	−0.302	0.297	−0.058	0.030	−0.065	0.104	−0.143	0.553	−0.235
		(8.08)	(−1.51)	(1.47)	(−0.30)	(0.16)	(−0.34)	(0.57)	(−1.18)	(3.62)	(−1.11)

Y	X	$\beta(3)$	$\beta(4)$	$\beta(5)$	$\beta(6)$	δ_1	δ_0	$\bar{R}^2$	SE of Est Adj	D-W	F-Statistic on All $F(6, 56)$
n_f	U_n	0.003	−0.007	0.009	−0.004	0.00003	−0.036	.864	0.004	2.143	2.886*
		(0.97)	(−2.62)	(3.54)	(−2.81)	(1.29)	(−0.98)				
U_n	n_f	−5.060	21.901	−32.547	24.041	−0.001	5.604	.916	0.351	2.097	2.229*
		(−0.29)	(1.31)	(−1.92)	(1.99)	(−0.63)	(1.76)				
n_f	m	0.262	−0.384	0.312	−0.103	−0.0002	−0.288	.853	0.004	2.031	1.983
		(1.47)	(−2.12)	(1.76)	(−1.02)	(−1.74)	(−2.31)				
m	n_f	0.041	0.075	−0.039	−0.137	−0.0001	−0.258	.999	0.006	2.000	0.309
		(0.13)	(0.25)	(−0.14)	(−0.69)	(−0.59)	(−1.25)				
n_f	rwp	−0.0001	0.0003	−0.0003	0.000001	0.00002	−0.045	.841	0.004	2.026	1.132
		(−0.51)	(1.06)	(−1.19)	(0.01)	(1.04)	(−1.18)				
rwp	n_f	91.625	128.847	−69.150	−53.598	−0.012	−37.639	.826	3.190	1.967	0.980
		(0.67)	(0.94)	(−0.50)	(−0.53)	(−0.66)	(−1.39)				
n_f	g	−0.004	0.020	−0.016	0.016	−0.000006	−0.073	.852	0.004	2.054	1.938
		(−0.24)	(1.38)	(−1.23)	2.09	(−0.11)	(−1.58)				
g	n_f	−0.555	−5.256	5.180	−2.354	0.002	1.066	.953	0.047	2.081	2.578*
		(−0.27)	(−2.62)	(2.44)	(−1.51)	(2.88)	(1.93)				
n_f	$g\$$	0.046	−0.105	0.073	−0.003	−0.0001	−0.113	.845	0.004	2.015	1.385
		(0.96)	(−2.16)	(1.46)	(−0.11)	(−1.23)	(−1.97)				
$g\$$	n_f	0.028	−0.746	−0.055	0.192	0.002	0.750	.999	0.015	1.692	1.343
		(0.04)	(−1.14)	(−0.09)	(0.41)	(3.86)	(3.14)				
n_f	p	0.160	−0.372	0.192	−0.033	−0.0002	−0.397	.840	0.004	2.060	1.072
		(0.74)	(−1.75)	(0.97)	(−0.26)	(−1.17)	(−1.58)				
p	n_f	−0.089	0.131	0.123	−0.223	0.0004	0.410	.9995	0.003	2.104	2.935*
		(−0.56)	(0.83)	(0.82)	(−1.86)	(2.38)	(2.10)				
n_f	w	0.010	−0.167	0.132	−0.060	−0.0004	−0.217	.841	0.004	2.002	1.125
		(0.06)	(−0.95)	(0.78)	(−0.56)	(−1.15)	(−2.00)				
w	n_f	−0.125	0.219	0.017	−0.053	0.0008	0.242	.9996	0.005	1.972	3.137**
		(−0.59)	(1.04)	(0.08)	(−0.32)	(2.25)	(2.25)				

NOTE.—t-statistics for coefficients appear in parentheses below relevant coefficients

*Significant at 5%

**Significant at 1%

All in all, the empirical results provide some evidence that the causal structure imposed on the data by the classical model of Section I is not obscenely at variance with the data. The evidence that m seems to be caused by RBaa means that the assumption that m is exogenous, embedded in the assumed autoregression of equation (1.6), must be abandoned. But this is not essential, since for the purpose for which the model is intended (unconditional forecasting), the regression in table 4 will do just as well Findings that contradict the model are that w seems to cause both RBaa and Un, according to both Sims's and Granger's tests. Also, according to Granger's test, m seems to cause Un, but according to Sims's test, it does not This last discrepancy requires reconciling, as does the apparently general tendency of Granger's test to reject exogeneity more readily than does Sims's test [19]

I do not believe that these results render a verdict on the model of Section I sufficiently negative for me to stop now before presenting estimates of the model The causal candidate that does the most damage to the hypotheses of the model is the money wage w, which does not appear itself as a variable in the model of Section I. Causal candidates drawn from the list of variables actually appearing in the model usually do not seem to violate the hypotheses of the model, which gives some encouragement to the project of estimating the model

IV. Estimates of the Model[20]

To estimate the model, a proxy for $E_{t-1}p_t$ was required.[21] As in a procedure previously used (Sargent 1973), the proxy for $E_{t-1}p_t$ was formed by regressing p_t against a list of variables dated $t - 1$ and earlier [22] In each case, this list included all the predetermined variables that appear on the right side of the equation in which $E_{t-1}p_t$ appears.

---

[19] In implementing Granger's test, I specified a maximal number of lagged own terms, usually four, upon which a variable was permitted to depend If the variable in question is exogenous but follows a mixed moving-average, autoregressive process so that its autoregression is of infinite order, this misspecification could lead to erroneous rejection of the hypothesis of exogeneity With Sims's test, premature truncation of the lag distribution will lead to too frequent rejection of the hypothesis of exogeneity when it is true (Christopher Sims points out to me that since the autoregressive part of the direct Granger regression whitens the residuals, thereby reseasonalizing them, it is not possible for the Granger test to "ignore" the seasonal bands, as the Sims test as applied here does This could conceivably account for some of the differences in the results of the two tests)

[20] The estimates of the model use the data seasonally adjusted by setting their Fourier transforms equal to zero in the seasonal bands, the same data used in the tests in Section III Estimates of the model using officially seasonally adjusted data were also made The results, which are qualitatively similar to those summarized here, are in the mimeographed appendix (n 17 above)

[21] For population (pop), I took the civilian population over 16 years old, while for the labor force I used the civilian labor force over 16. The labor force participation rate ηf was measured as the ratio of the latter to the former The total civilian unemployment rate was used Notice that $\eta f_t + pop_t - Un_t$ approximately equals civilian employment, so my production function views GNP as a function only of civilian employment.

[22] To form the proxy for $E_{t-1}p_t$, p_t was regressed on a constant, trend, three seasonal dummies, and p, w, ηf, and Un, each lagged one through four times.

Since the model is a simultaneous one, an instrumental-variables estimator was used to estimate the coefficients Current endogenous variables that appear on the right side of an equation were replaced by the systematic part of a regression of that variable on the same variables that were used to form the proxy for $E_{t-1}p_t$ plus current values of the exogenous variables.[23]

The estimates are reported in table 9 The production function includes current and four lagged values of $n_t \equiv (n_{ft} + pop_t - Un_t)$ The estimates of the production function (item 3) are compatible with increasing returns to labor in the short run and slightly decreasing returns to labor in the long run

The estimates reported in table 9 possess signs that agree with a priori expectations Unexpected increases in the price level are estimated to increase the labor force participation rate and decrease the unemployment rate Increases in the unemployment rate decrease the labor force participation rate, which is consistent with a discouraged-worker effect.

In the estimates reported in table 9, I have not included innovations in Z_t as determinants of R, so that the equation for R (the Baa rate) is simply an autoregression Two pairs of equations for portfolio equilibrium are reported The first pair regresses the reciprocal of the log of velocity $(m - p - y)$ against current- and lagged-interest rates, one member including and the other excluding trend Including trend is seen to increase the coefficients on current- and lagged-interest rates and to make their sum positive This is a common, though widely ignored, result including a trend in postwar estimates of demand schedules for money for the United States tends to eliminate any inverse dependence of velocity on interest rates The second pair of portfolio balance equations regresses $m - p$ on current and lagged y's and R's, again with and without trend Including trend again has important effects on the coefficients For my purposes, any of these four or any other reasonable demand schedule for money is suitable Notice also that the model will work in the same "recursive" way if a demand schedule for money is dropped and replaced by a regression of $p_t + y_t$ on current and lagged m, the sort of equation estimated by Sims (1972*b*) and Andersen and Carlson (1970)

---

[23] The endogenous variables were replaced by the systematic part of a regression of themselves against pop_t, m_t, g_t, $surp_t$, the log of current government employment, and all of the variables reported in n 22 above. The reader may wonder whether eqq. (1) and (2), which have lagged endogenous variables as regressors, can be consistently estimated by the technique employed If the residuals are serially correlated, my estimates are not consistent But it is straightforward to show that, e g , the *Un* vs p exogeneity tests of Section III can be viewed as tests for serial correlation of the disturbances in eq (1), failure to reject exogeneity of unemployment (p's failing to cause *Un*) being consistent with no serial correlation. In effect, then, some testing for the null hypothesis of no serial correlation has been carried out, with results favorable to the null hypothesis

TABLE 9

ESTIMATES OF THE MODEL (1951 I–1973 III)

Variable	Estimate
1 Un_t	$\underset{(2.0)}{-0.287}(\hat{p}_t - E_{t-1}p_t) + \underset{(2.5)}{0.0043} + \underset{(0.5)}{0.0000007t} + \underset{(12.8)}{1.47}Un_{t-1}$ $- \underset{(2.9)}{0.59}Un_{t-2} - \underset{(0.1)}{0.03}Un_{t-3} + \underset{(0.3)}{0.04}Un_{t-4}$*
2 nf_t	$\underset{(0.9)}{0.149}(\hat{p}_t - E_{t-1}p_t) - \underset{(1.9)}{0.075}\hat{U}n_t - \underset{(1.3)}{0.038} + \underset{(2.1)}{0.00004t}$ $+ \underset{(8.2)}{0.94}nf_{t-1} - \underset{(0.7)}{0.11}nf_{t-2} - \underset{(0.2)}{0.02}nf_{t-3} + \underset{(1.0)}{0.12}nf_{t-4}$†
3 y_t	$\underset{(3.5)}{1.09}\hat{n}_t + \underset{(1.0)}{0.24}n_{t-1} - \underset{(1.0)}{0.24}n_{t-2} - \underset{(0.6)}{0.14}n_{t-3} - \underset{(0.1)}{0.02}n_{t-4}$ $+ \underset{(1.8)}{0.35} + \underset{(4.5)}{0.0009t}$‡
4 R_t	$\underset{(13.1)}{1.52}R_{t-1} - \underset{(3.7)}{0.77}R_{t-2} + \underset{(2.4)}{0.44}R_{t-3} - \underset{(2.1)}{0.24}R_{t-4} + \underset{(1.8)}{0.15} + \underset{(2.0)}{0.0034t}$§
5a $m_t - p_t - y_t$	$-\underset{(0.0)}{0.0004}\hat{R}_t - \underset{(0.0)}{0.0004}R_{t-1} - \underset{(1.4)}{0.010}R_{t-2} + \underset{(2.4)}{0.021}R_{t-3}$ $+ \underset{(0.9)}{0.007}R_{t-4} + \underset{(2.1)}{0.015}R_{t-5} - \underset{(1.6)}{0.012}R_{t-6} + \underset{(1.0)}{0.007}R_{t-7} - \underset{(212.0)}{0.91}$ $- \underset{(14.7)}{1.37} \times 10^{-3} \times t$‖
5b $m_t - p_t - y_t$	$-\underset{(2.0)}{0.032}\hat{R}_t - \underset{(0.7)}{0.006}R_{t-1} - \underset{(3.2)}{0.027}R_{t-2} + \underset{(1.6)}{0.014}R_{t-3} - \underset{(0.9)}{0.007}R_{t-4}$ $- \underset{(0.0)}{0.0003}R_{t-5} - \underset{(2.0)}{0.018}R_{t-6} - \underset{(0.5)}{0.004}R_{t-7} - \underset{(59.2)}{0.22}$#
5c $m_t - p_t$	$-\underset{(0.5)}{0.0060}\hat{R}_t - \underset{(1.2)}{0.0059}R_{t-1} - \underset{(1.7)}{0.0091}R_{t-2} + \underset{(2.6)}{0.0143}R_{t-3}$ $+ \underset{(1.6)}{0.0080}R_{t-4} + \underset{(2.0)}{0.0107}R_{t-5} - \underset{(0.4)}{0.0022}R_{t-6} + \underset{(0.8)}{0.0042}R_{t-7}$ $+ \underset{(3.3)}{0.45}\hat{y}_t + \underset{(1.9)}{0.16}y_{t-1} + \underset{(2.3)}{0.19}y_{t-2} + \underset{(1.1)}{0.09}y_{t-3} - \underset{(0.9)}{0.06}y_{t-4}$ $+ \underset{(0.3)}{0.02}y_{t-5} + \underset{(0.8)}{0.05}y_{t-6} + \underset{(1.1)}{0.06}y_{t-7} - \underset{(2.6)}{0.22} - \underset{(2.1)}{0.0003t}$**
5d $m_t - p_t$	$-\underset{(0.2)}{0.0023}\hat{R}_t - \underset{(1.5)}{0.0075}R_{t-1} - \underset{(1.6)}{0.0089}R_{t-2} + \underset{(1.7)}{0.0088}R_{t-3}$ $+ \underset{(0.9)}{0.0045}R_{t-4} + \underset{(0.9)}{0.0043}R_{t-5} - \underset{(0.8)}{0.0044}R_{t-6} + \underset{(0.1)}{0.0003}R_{t-7}$ $+ \underset{(2.5)}{0.25}\hat{y}_t + \underset{(0.9)}{0.06}y_{t-1} + \underset{(1.3)}{0.09}y_{t-2} - \underset{(0.1)}{0.01}y_{t-3} - \underset{(2.2)}{0.14}y_{t-4}$ $- \underset{(0.3)}{0.02}y_{t-5} + \underset{(0.2)}{0.01}y_{t-6} + \underset{(0.6)}{0.04}y_{t-7} - \underset{(1.7)}{0.052}$††

NOTE —t-statistics are in parentheses beneath coefficients Hatted variables (^) are systematic parts of regressions against instrumental variables

* $\bar{R}^2$ = 908, SE = 371 D-W = 2.07

† $\bar{R}^2$ = 867, SE = 0040 D-W = 2.00

‡ $\bar{R}^2$ = 95 SE = 00964, D-W = 2.03, filter $(1 - 6L)^2$, sum of weights on n = + 93, $\hat{n}_t \equiv (nf_t - \hat{U}n_t + pop_t)$, $n_t \equiv (nf_t - Un_t + pop_t)$

§ $\bar{R}^2$ = 99, SE = 158 D W = 1.89, sum of weights = 95

‖ $\bar{R}^2$ = 93 SE = 00856, D–W = 1.97, filter $(1 - 6L)^2$, sum of coefficients on R = + 024

# $\bar{R}^2$ = 28 SE = 01085 D–W = 1.91, filter $(1 - 8L)^2$, sum of coefficients on R = − 080

** $\bar{R}^2$ = 29, SE = 00561, D–W = 2.10, filter $(1 - 8L)^2$, sum of weights on R = + 014, sum of weights on y = + 96

†† $\bar{R}^2$ = 25, SE = 00576, D–W = 1.99, filter $(1 - 8L)^2$, sum of weights on R = − 005, sum of weights on y = + 28

V. Conclusions

This paper has estimated and tested a macroeconometric model with "classical" or "monetarist" policy implications, even though it has "Keynesian" short-run properties. Some evidence for rejecting the model has been turned up, but it is far from being overwhelming and decisive The evidence that seems most damaging to the model comes from the role that the money wage plays in apparently "causing" unemployment and the long-term interest rate On the other hand, the tests have turned up little evidence requiring us to reject the key hypothesis of the model that government monetary and fiscal-policy variables do not cause unemployment or the interest rate The fact that such evidence has been hard to turn up ought to be disconcerting to users of the existing macroeconometric models, since as usually manipulated those models all imply that monetary and fiscal policy *do* help cause unemployment and the interest rate

Models of the kind presented in this paper imply that there is no scope for the government to engage in activist countercyclical policy, so that it might as well employ rules without feedback for fiscal and monetary policy, for example, Friedman's *x* percent growth rule for the money supply In contradistinction, macroeconometric models as they are usually manipulated imply that it is optimal for the government to use rules with feedback, which may imply "leaning against the wind," contrary to Friedman's rule If we are to have any reason to believe that rules with feedback are superior to rules without feedback, there should be empirical evidence in hand that some existing macroeconometric model can outperform models of the class studied in this paper It is my impression that such evidence does not yet exist

References

Andersen, Leonall C., and Carlson, Keith M "A Monetarist Model for Economic Stabilization." Federal Reserve Bank of St Louis *Review* (April 1970)

Bodkin, Ronald G , and Klein, Lawrence R "Nonlinear Estimation of Aggregate Production Functions " *Rev. Econ and Statis* 49 (February 1967) 28–44

Cooper, J Phillip, and Fischer, Stanley "Monetary and Fiscal Policy in the Fully Stochastic St Louis Econometric Model " *J Money, Credit, and Banking* 6 (February 1974) 1–27

Granger, C W J "Investigating Causal Relations by Econometric Models and Cross-spectral Methods " *Econometrica* 37 (July 1969) 424–38

Lucas, Robert E , Jr "Capacity, Overtime, and Empirical Production Functions " *A.E.R Papers and Proc.* 60 (May 1970) 23–27

——— "Econometric Testing of the Natural Rate Hypothesis " In *The Econometrics of Price Determination Conference*, edited by Otto Eckstein Washington Board of Governors, Federal Reserve System, 1972. (*a*)

——— "Expectations and the Neutrality of Money." *J Econ Theory* 4 (April 1972) 103–24 (*b*)

——— "Some International Evidence on Output-Inflation Trade Offs." *A.E.R* 63 (June 1973) 326–34

Modigliani, Franco, and Shiller, Robert. "Inflation, Rational Expectations, and the Term Structure of Interest Rates " *Economica*, n s 40 (February 1973) 12–43

Poole, William. "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model." *Q.J.E.* 84 (May 1970): 197–216.

Sargent, Thomas J. "A Note on the Accelerationist Controversy." *J Money, Credit, and Banking* 3 (August 1971). 721–25

———. "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment." In *Brookings Papers on Economic Activity*, vol. 2, edited by Arthur M Okun and George Perry Washington. Brookings Inst., 1973

Sargent, Thomas J, and Wallace, Neil. "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J P E.* 83, no 2 (April 1975) 241–54

Shiller, Robert "Rational Expectations and the Structure of Interest Rates." Ph D dissertation, Massachusetts Inst Tech, 1972

Sims, Christopher A "The Role of Approximate Prior Restrictions in Distributed Lag Estimation." *J American Statis Assoc* 67 (March 1972) 169–75 (*a*)

——— "Money, Income, and Causality" *A E R* 62 (September 1972) 540–52 (*b*)

——— "Seasonality in Regression" *J American Statis Assoc* 69 (September 1974) 618–26

Tobin, James "Inflation and Unemployment" *A E.R* 62 (March 1972) 1–18.

Wachter, Michael "A Labor Supply Model for Secondary Workers" *Rev. Econ and Statis* 54 (May 1972) 141–51

Wallis, Kenneth S "Seasonal Adjustment and Relations between Variables" *J American Statis Assoc* 69 (March 1974) 18–31

Economic Development in Communist China

Colin Clark

Monash University

This text is an extension of a previous study, covering the period from the 1930s to 1959, published by the present writer in 1965. The results of the earlier study (amended) are included in the final table. Economic progress was violently interrupted by "The Year of the Great Leap Forward" (1958), when, as a result of misinformation about agricultural labor requirements and hysterically falsified statistics, it was claimed that agricultural output was being doubled in 1 year, and that immense transfers of labor to other employments were immediately possible. The result was acute agricultural shortages, indeed famine in 1960–61, and complete disruption of industrial production. Recovery from these disasters took several years. A lesser interruption (to industry but not to agriculture) took place in the "Cultural Revolution" of 1966–67. Those to whom the idea of a labor shortage in China appears paradoxical must be reminded that China has few draft animals and still fewer tractors. To cultivate a country the size of China with hand hoes requires several hundred million workers. Chinese population is probably substantially lower than is generally believed, and almost certainly has not been expanding at the rate of 2 percent per year frequently attributed to it. Famine conditions in the early 1960s caused a considerable reduction in the rate of population growth. Publication of Chinese official statistics virtually ceased in 1959, and sources of information for subsequent years are indirect and complex. The basis of the methods used is the construction of estimates of agricultural output and industrial production index numbers, supplemented by information about employment and wages. Almost all attempts hitherto to state China's (and other developing countries') national product in dollar terms give results considerably too low. The yuan has a high purchasing power over services and over some labor-intensive commodities. For a true comparison all Chinese consumption of food should also be revalued at U.S. retail prices. On this basis Chinese 1971 gross product per head, expressed in U.S. dollars of 1974 purchasing power, was $154 for food, $140 for other private consumption, and $157 for investment and government services, or $451 in all. The long-run rate of increase of real gross product per head of population

[*Journal of Political Economy*, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved.

has been about 2 percent per year, whether we take the 1930s or the early 1950s as our starting point. This rate is a little below the general average for developing countries, and much less than is usually claimed. Inequalities in income distribution in China are also not very different from those prevailing in other countries

The valuation of China's national product in dollar terms was pioneered by Hollister (1958), by Liu and Yeh (1963), and by Ishikawa (1965) A previous study by the present writer (Clark 1965) made a valuation up to 1959 and serves as the basis for the present study

As is well known, it is far more difficult to obtain information for the years since 1959 than for the previous period

Chairman Mao had bad advice on agricultural economics He believed that about one-third of the agricultural labor force of China was unoccupied (Mao)—"under present conditions of production there would be a one-third man power surplus in the rural areas"—and could therefore be transferred to other employments without reducing agricultural output Who gave him this misinformation nobody knows—perhaps advisers from other Asian countries with different climatic conditions and methods of cultivation So far from having a vast reserve of unoccupied labor, Chinese agriculture, except in midwinter, was subject to labor shortages This was made perfectly clear in Buck's surveys (1964) in the 1930s, as well as from other sources of information

To those to whom the idea of labor shortage in China seems paradoxical, it must be pointed out that China, even now, has hardly any tractors and few draft animals If you take a country the size of China and have to plan for its cultivation with hand hoes, you are going to need several hundred million workers

Rarely has the ruler of a country been so grossly misinformed, and probably never with such utterly disastrous consequences. For in 1958 Mao decreed "The Year of the Great Leap Forward," in which the whole farm population was forcibly grouped into communes According to contemporary reports, families were to be broken up, to live separately in communal dormitories, and only to meet each other once a month This program may indeed have been intended but was not carried out Farm labor was diverted, apparently to the extent of 100 million or more, to industrial employment, to the construction of dams and roads (mostly by unaided man power, without mechanical equipment), and also to fantastic projects for producing steel and cement in small backyard furnaces.

The chaos was compounded by the complete breakdown of the Chinese statistical system, which has never since recovered There has in fact been an almost complete suppression of statistical information from China ever since, and estimates have had to be prepared by indirect methods (table 10)

The suppression of all official statistical information does not prove, but certainly indicates, that economic results have been disappointing. The principal sources of information (as will be seen from the references) are *Current Scene*, published by the U S. Consulate-General in Hong Kong; publications of two private Hong Kong research organizations, *China News Analysis* and *Union Research* (publisher of *Communist China*), *China Quarterly* (published in London), and also material assembled by Swamy (1973) and by "Economic Indicators for China" (1974)

Chinese officials now recognize Chairman Mao's mistake, though his mistakes are still shared by some foreign economists. "I think that the marginal product of a worker in agriculture is less than his consumption," wrote Tobin (1974) "Our Chinese hosts disagree . they repeatedly claimed that there was a shortage of labour in agriculture "

The Chinese population is not known within wide limits. A census taken in June 1953 indicated a population (excluding Taiwan) of 583 million This figure was accepted at the time by most (though not all) of the outside critics who have studied the available evidence. Subsequently, however, more doubts have been expressed

The official demographic adviser to the U S State Department, whose figures are quoted in the report on China prepared by the Joint Economic Committee (U S Congress Joint Economic Committee 1968), accepted a statement made in some Chinese official publications in the 1950s (though not repeated subsequently) to the effect that the Chinese population was growing at a rate of over 2 percent per year, and he has gone on woodenly adding these percentages to the original census figure, to end up with over 900 million for 1974 Similar high figures are given by Swamy (1973) and "Economic Indicators for China" (1974) But they must be disputed. If the population were indeed as large as this, the food situation would be far worse than is described below

An estimate (Clark 1966) making use of all the available information on deaths, births, and the age composition of the population indicated an average rate of increase between 1953 and 1958 of 1 6 percent per year. In the troubled years 1958–63 it was estimated (admittedly on very little evidence) that the rate of increase fell to as low as 0 4 percent per year, recovering to 1 1 percent per year after 1963

The official *Chinese Atlas* in 1957 estimated the growth rate at only 1 3 percent (birth rate 25, death rate 12 per thousand) The experienced demographer Spengler (1962–63), on the other hand, referring to the early 1960s, estimated the average completed Chinese family at 6 2 (Buck [1964] estimated about 5 5), giving a crude birth rate of 40 per thousand but a death rate of 32.5, or a rate of increase of only 0.75 percent per year

A leading authority on Chinese population, L A. Orleans (1969, 1972, 1974), agrees with 1 6 percent per year for 1953–58 but estimates much

TABLE 1

CHINA'S POPULATION

A. POPULATION INCREASE (% PER YEAR)

	Orleans	Clark	Compromise
1953–58	1.6	1.6	1.6
1958–65	1.4	0.4	0.9
1963–70	1.4	1.1	1.25
1970–75	1.3	1.1	1.2

B. MIDYEAR POPULATION ESTIMATE*

	Population (in Millions)
1953	583
1958	635
1963	664
1970	724
1975	768

* Based on 1953 census and compromise rates of increase.

less decline in the rate of growth in 1958–63. Since 1963, falls in the mortality rate, he estimates, have been matched by a decline in births. There is certainly evidence that there has been an active campaign for family limitation, first in the cities (partly perhaps in view of an acute shortage of housing) but now in the countryside also (table 1).

Higher figures have certainly been quoted, but there has been a general tendency to revise them downward (see table 1). "For five years [after 1957] official Chinese statements used a constant figure of 650 million. This was raised to 700 and eventually to 750 million, but the editorial published in the *People's Daily* on New Year's Day 1970 reverts to quoting a figure of 700 million" (Klatt, private communication). Field (1970*a*) prepared estimates for each province and large city for 1970, totaling 729 million, very close to the above estimate (he also quotes the official estimates for 1957, totaling 650 million, but points out that these were based on assumed growth rates since 1953 averaging 2 percent per year, which was too high).

Some kind of national register has always been in existence, and there may have been an attempt to bring it accurately up to date for use as a basis for population estimation in 1964, though Klatt (private communication) considers that such action was limited to a few localities. "On 27th August 1972 Premier Chou En-lai told us that a census was taken in 1964–5 recording around 700 million, with natural increase 2% per year" (Chen 1973). Possibly he meant a national estimate based on sample surveys. "Brigades" visited in rural China in 1972 averaged a rate of natural increase of 1.8 percent per year, as against 1.2 percent in Peking and 0.7 percent in Shanghai in 1971 (Brown 1973; Chen 1973).

Field (1970*a*) considers that these estimates refer to mid-1964, and gives provincial and principal city estimates totaling 711 million for that date

However, Chairman Mao told his old friend Edgar Snow that the population in 1964 was "much below 680 million" (Snow 1965)

"The Chinese Government itself is without population data," writes Orleans (1972, p 34) This was shown by a very frank interview given by Vice-Premier Li Hsien-Nien to a correspondent of a Cairo newspaper in November 1971, in which he states that different government departments had given him estimates ranging from 750 to 830 million "However the Planning Department insists that the number is less than 750 million" (Chen 1973)

It is difficult to see how the Planning Department can do any planning if it does not even know the size of the population for which it has to provide

The same vice-premier gave an estimate of 750 million for 1970 to a party of Japanese visitors (British Broadcasting Corporation 1972)

Calorie requirements vary with climate, body weight, the amount of hard physical work done, and the proportion of children in the population (Clark and Haswell 1964) In Southeast Asia, with small-bodied populations, in a hot climate, not engaged in hard physical work throughout the year, calorie requirements may be as low as 1,650 per person per day In northern China, on the other hand, where each of the above conditions is reversed, the figure rises to about 2,000 For China as a whole we may take a figure of 1,900

In comparing these figures with calories actually available, we must take into account a limited amount of unavoidable wastage in food storage, preparation, etc

Buck (1965) made a careful comparison, taking full account of crops used for seed, fodder, and other purposes, and not available for human consumption, between his own previous results for 1929–33 and the information for 1949–58 Even in this comparatively favorable period, the average Chinese was much worse fed than before the Revolution Average calories had fallen from 2,410 to 2,107

If we accept the official Chinese figures, which show harvests at a low level in 1949, rising progressively through the 1950s, we have to deduce (Liu and Yeh 1961) that the Chinese were living on 1,190 calories per head in 1949, at which level, if continued for any length of time, they would certainly have died Understating past figures in order to show a spuriously high rate of growth subsequently is one of the favorite devices in that well-known textbook *How to Lie with Statistics*. The official figures rose gradually but still stood at only 1,830 in 1957, rising to a sudden peak of 2,440 in 1958. Klatt (private communication) estimates (after adjusting his population figures to those given above) as follows: for 1952–53, 1,950 calories per head per day; 1957–58, 2,000; 1965–66,

TABLE 2

AGRICULTURAL PRODUCTION (MILLIONS OF TONS), 1958–65

	1958	1959	1960	1961	1962	1963	1964	1965
Grain*	204	170	160	165	178	183	195	193
Cotton	2.1	2.4	1.4	1.1	1.1	1.2	1.2	1.2
Soya beans	10.1†		5.7	5.5		7.8	6.9	6.8
Groundnuts	2.6†					1.9	2.3	
Sugar	0.9	1.1	0.9	0.7	0.5	0.5	1.1	1.5
	Swamy's (1973) Estimates‡							
Soya beans	10.5	11.5		5.5	6	6.5	7	7.5
Fibers	2.10	2.41		1.10	1.05	1.22	1.61	1.85
Oil seeds	4.24			2.3				3.4
Other nonfood				5.93				11.7

* Including grain equivalent of root crops.

† 1957.

‡ For noncereal crops, Swamy (1973) quotes the official figures for the 1950s and also gives the estimates shown in the table.

2,100, 1970–71, 2,225. Even with the recent improvements, however, food consumption standards are still far below those of 1929–33.

In the hungry period which followed "The Year of the Great Leap Forward," food supplies fell to famine levels (Clark 1965) of 1,800 calories per day or less. The worst cases of refugees reaching Hong Kong in the preharvest hunger period in 1961 reported their average family consumption per head as 1,150 for urban families and only 900 for rural (privately communicated results, which in the first place were per adult multiplied by a factor of 0.7).

The loss of agricultural production caused by "The Year of the Great Leap Forward" was catastrophic. Even by 1965, the 1958 output had not been recovered—by which time population had presumably substantially increased (*Current Scene*, various issues, and contemporary reports) (table 2).

Since 1965, the improvement in agricultural production has been at a rate higher than the rate of population growth (*Current Scene*, various issues). The "Cultural Revolution," which severely damaged industrial productivity, had little effect on the countryside (table 3).

According to Swamy (1973), Chou En-lai informed Edgar Snow, in a personal interview, that in 1971 China had a grain reserve stock of 40 million tons. "Economic Indicators for China" (1974) gives grain output throughout this period as some 10 percent higher—perhaps to match that publication's higher population estimates. Its estimates for cotton are also 300,000–500,000 tons higher.

In comparing these figures with the improvement in per head calorie supply at the rate of a little over 1 percent per year estimated by Klatt

TABLE 3

AGRICULTURAL PRODUCTION (MILLIONS OF TONS), 1965–73

	1965	1966	1967	1968	1969	1970	1971	1972	1973
Grain*	193	190	200	205	205	225	225	215	228
Cotton	1.2	1 6		1.7	1 6	1 7	1.7	1 4	
Soya beans	6 8	6 8	6 9	6 5	6 2	6.9	6 7	6 3	6.7
Groundnuts			2 3	2 2	2 3	2.6	2 6	2 4	2.6
Rapeseed			0.8	0 8	0 7	0 8	0.8	1 0	1.1
Sugar	1 5	1 6	1 7	1 8					

* Including grain equivalent of root crops

(see above), it must be remembered that a proportion of the output of agriculture is required for industrial raw materials and for exports

One of the principal factors in this improvement was a greatly increased use of chemical fertilizers This will be unwelcome news for some admirers of China, namely, the cranks who believe that agriculture needs only "organic fertilizers" and commend the Chinese as examples to the world because of the large quantities of night soil, duckweed, pond mud, etc., which they spread on the soil with such great labor The Chinese also supplemented their own production of chemical fertilizers by using some of their limited supply of foreign exchange to import all that they could, principally from Japan (table 4)

Chinese fertilizer policy has been much more rational than that of India, which botched both production and import of this vital requirement India consequently is experiencing only slowly rising agricultural output and continued hunger Swamy (1973) admits this point regarding fertilizers (and also considers that China has made more efficient use of its capital resources and of the possibilities of handicraft production than has India, even though his general conclusion is that India's economic growth per head of population has been more rapid than China's over the last two decades)

There is unfortunately considerable diversity in the estimates of the amounts produced and used Field's figures (1970*b*) of production, which are much lower than Rawski's (1973*b*), apparently refer to nitrogenous fertilizers only

Also, unfortunately, we lack knowledge of the composition of the fertilizer supply (nitrogenous, phosphatic, and potassic) and hence of its expected marginal productivity A kilogram of pure nitrogen is expected to yield 10 kilograms additional output of grain, a kilogram of pure phosphorus pentoxide, little less The principal nitrogenous fertilizer used, ammonium sulphate, contains only 20 percent nitrogen, but urea (now increasingly used) contains 46 percent, and superphosphate, the principal phosphatic fertilizer, contains a little over one-third phosphorus pentoxide Swamy (1973) estimates the marginal return in grain for the

TABLE 4

CHEMICAL FERTILIZERS (MILLIONS OF METRIC TONS)

	PRODUCTION		SUPPLY (INCLUDING IMPORTS)		
YEAR	"Economic Indicators for China" (1974)	Rawski (1973*b*)	"Economic Indicators for China" (1974)	Chao (1970)	Swamy (1973)
1952	0.2		0.4		0.32
1953					0.59
1954					0.80
1955					1.26
1956					1.61
1957	0.8	0.9	1.9	1.9	1.94
1958	1.4	1.2	3.0	2.7	2.72
1959	1.9		3.1		2.93
1960	2.5	2.5	3.5	3.4	3.67
1961	1.5		2.5		3.54
1962	2.1	3.0	3.1	4.0	4.22
1963	3.1	4.2	5.1	6.0	5.11
1964	4.2	5.9	5.4	7.0	6.00
1965	5.7	8.9	8.0	10.9	
1966	7.4	11.6	9.9	13.9	
1967	5.9		10.2		
1968	8.2		12.2		
1969	11.3		15.4		
1970	14.0	14.0	18.3		
1971	16.9	16.8	21.2		
1972	19.9		24.1		
1973	24.8		28.9		

whole fertilizer input at five, but three would probably be a more correct figure.

For the later 1960s, as against approximately 4 million tons annually of imports, Rawski (1973*b*) estimates 8 million, costing US$200 million per year. Judging from the price, this must have consisted mainly of phosphatic rather than nitrogenous fertilizers.

The most recent information (*Nihon keizai shimbun* 1975) indicates that imports from Japan alone for the half-year to June 1975 will include 600,000 tons of urea and 166,000 tons of ammonium sulphate, that is, 312,000 tons of pure nitrogen equivalent at a price exceeding $160 million.

On this basis we might have expected, between 1957 and 1973, a rise in output, due to fertilizer applications alone, of as much as 81 million tons of grain. The increased output of grain over this period was estimated above at only 43 million tons (65 million tons on the "Economic Indicators for China" [1974] estimates). The increased output of other crops, converted to economic grain equivalents, probably did not exceed 10 million tons. It does not appear that the efficiency of the Chinese in the application of fertilizers has been as great as Swamy gave them credit for.

There has been another important factor in the recovery of Chinese agriculture, which will also give pain to many of China's admirers, namely, the restoration of private plots, which were conceded as the excesses of the Cultural Revolution died down. Their average sizes have been given (Walker 1965) in square meters per household as 483 square meters per household in northern China, 154 in western and central China, and 87 in southern China These figures related to the mid 1950s Since then, they were probably reduced in "The Year of the Great Leap Forward" but subsequently raised again

These differences in average size are to some degree offset by the warmer and more humid climate of the south, where multiple cropping is possible The private plots are largely used for growing vegetables, at yields which Walker estimates at 37 tons per hectare Even allowing for a 75–80 percent water content in vegetables, these yields are much higher than the 3 tons of rice or 1 ton of wheat obtainable per hectare in farming These small private plots—as in Russia—are now estimated ("Private Plots" 1971) to be providing 25 percent of the entire agricultural output

Another important concession to free-market principles has been that, in place of universal price planning, "provincial and local commerce departments are now empowered to vary prices according to product quality, supply and demand" ("PRC Economy in 1973" 1974)

The number of tractors is gradually increasing, and they sufficed for 8 percent of the cultivated land in 1965 Swamy (1973) gives the following estimates, in millions of horsepower for 1952, 0 03, 1953, 0 04, 1954, 0 08, 1955, 0 12, 1956, 0 30, 1957, 0 56, 1958, 0 69, 1959, 0 90, 1960, 1 22, 1961, 2 23, 1962, 3 6, 1963, 5 2, 1964, 7 0

Kang Chao (1970) confirms a serious loss of cultivated land (through mistaken irrigation works, etc) from 113 million hectares in 1957 to 102 million in 1963–65 About 40 percent of all land was double-cropped, but a similar proportion had been double-cropped in the 1950s

Meat, eggs, and fish play only a small part in the Chinese diet, and dairy products none at all It is said that, for some unknown physiological reason, the southern (though not the northern) Chinese (together with some Southeast Asian peoples) lack the enzyme lactase (except in infancy) and so have difficulty in digesting milk and dairy products (Liu and Yeh 1963, Clark 1965, Klatt 1967) (It is said, however, that Chinese lose this disability when they emigrate and that it may be environmental rather than genetic.) Meat, fish, and eggs (valued at international prices) constituted 13 percent of food production in 1929–37, 17 percent in 1952, and 19 percent in 1957

The number of draft animals reached a peak (State Statistical Bureau 1960) in 1956 at 66.0 million cattle and buffalos, 7 4 million horses, and 13 4 million mules and donkeys, or 86.8 million in all, but was already declining by 1959. Draft animals, it must be remembered, will not and cannot work unless well fed, which means that they have to compete with humans for the limited supplies of grain and root crops because there is little grazing or fodder crop cultivation in China So in times of food shortage they are of little value and often have to be slaughtered Ishikawa (1965) considers that the total number of draft animals was only 50–60 million in 1965, though Klatt (1966) thinks that by then the 1957 cattle numbers had been recovered The number of pigs, which provided about 60 percent of the meat supply in the 1950s, rose (State Statistical Bureau 1960) from 90 million in 1952 to 146 million in 1957 In 1965 the *Soviet Encyclopaedia* estimated their numbers again at 150 million, and Wenmohs makes a similar estimate (1967) Klatt and Ishikawa raise the figure to 180 million Wenmohs considers that after 1965 pig numbers declined but other livestock increased The numbers of sheep and goats in 1967 were still much below their 1959 level of 112 5 million (Wenmohs 1967)

The evidence about livestock products is very inadequate, but it seems unlikely that there has been any significant improvement in per head supplies since the 1950s

The available information on the output of industry and construction calls for an analysis of some complexity

Figures of industrial production, usually stated in yuan of a constant 1952 purchasing power, are confused by the Chinese custom (borrowed from the Russians) of including all output gross If we include yarn and cloth, coal, pig iron, steel, and machinery, all at their gross values, we will have been guilty of a great deal of duplication (or triplication or quadruplication) of the values of some products. Nor is it safe to assume that the degree of duplication is approximately constant, so that the duplicated values at any rate can be used to give an index of true net output—as industry develops and becomes more complex, the degree of duplication increases Thus Rawski (1973*a*), using a 1952 base of 100 gross output, quotes a figure of 241 for 1957, but Kang Chao estimates that, on a properly constructed index number (Chao 1963) weighted by net outputs, it rose only to 186 Rawski, however, acquits the official figures of the charge of another, more subtle statistical offence (which was at one time widely practiced in Russia), namely, entering new, and rapidly growing, industries at anomalously high prices, so as to overstate their relative contribution to the total.

Rawski was able to obtain figures for all industrial production in Shanghai (some 20 percent of all China), which he considers indicative of production in China as a whole These are compared with properly compiled index numbers (Field 1970*b*; Cheng 1973) (see table 5)

Perkins (1967, 1971) estimates net output free of duplication, also in yuan of 1952 purchasing power. His totals agree with other index numbers. For 3 years he also provides a subdivision into classes of goods. Rawski also gives indexes (some of them from provincial sources) of the output of consumption goods, again in gross terms. However, if we make

TABLE 5

INDUSTRIAL PRODUCTION

	1957	1965	1970	1971
Shanghai production (billions of 1952 yuan) .	12.8	27.3	43.6	47.5
Above expressed as index number	100	213	340	371
Field-Cheng index number	100	153	230	253

TABLE 6

ANALYSIS OF INDUSTRIAL OUTPUT

	1952	1957	1965	1966	1970	1971
Net output in billions of 1952 yuan (Perkins 1967)						
"Producers' goods"	3.15	11.07	22			
Minerals	1.47	3.07	5			
Consumption goods	3.30	4.92	10			
Handicrafts	4.72	5.38				
Total	12.6	24.5	37	41		
Above expressed as index number (1957 = 100)	53	100	151	167		
Chao-Field-Cheng index	58	100	153	167	229	252
Ratio of consumption goods to all production, 1957 base (Rawski 1973*a*, 1973*b*)		1.00	0.69		0.86	0.90
Index of consumption goods output (product of last two lines)		100	106		197	227

the perhaps less impermissible assumption that the degrees of duplication in consumption-goods output and in general production are not necessarily the same, but only that they are rising at the same rate, we can obtain the change in the ratio of consumption goods to total output (see table 6).

So calculated, the consumption-goods output for 1965 shows a rise of 6 percent over 1957, as compared with Perkins's fall of 3 percent. Perkins's rise of 28 percent between 1952 and 1957 may be compared with a rise of 25 percent in consumption other than food and rent over this period, estimated by me by an entirely different method (Clark 1965).

The industrial production index includes the output of certain construction materials (cement, timber, structural steel), but it is convenient to include at this stage the whole product of the construction industry.

Between 1952 and 1957 cement output (priced at 85 yuan per ton) rose from 0.24 to 0.56 billion yuan, and timber output (Stanford Research

Institute) from 1.22 and 1.95, or from 1.46 to 2.53 billion 1952 yuan taken together.

Gross values of construction were estimated as shown below (billion yuan at current prices).

	1952	1953	1954	1955	1956	1957
Hollister (1958)	5.45					
Liu and Yeh (1963)	5.17	7.84	8.67	9.48	15.18	12.31
Ishikawa (1965)	3.71	6.51	7.50	8.63	13.99	12.64

We must reduce the 1957 figure to allow for a rise in prices since 1952 (which Hollister estimates at 11.4 percent), the results, in billions of 1952 yuan, are shown below.

	1952	1957
Cement and timber output	1.46	2.53
Gross value of construction	5.45	11.3
Ratio	3.7	4.4

This high and rising ratio represents (1) wages of construction workers, which we would expect to be relatively low in China, (2) locally produced bricks, tiles, thatch, etc., not included in the statistics of industrial output, (3) other industrial products included in construction (glass, nails, paint, etc.). In estimating the amount to be *added* to the industrial production figures in order to include all construction, we must take account of items 1 and 2 but not 3.

In 1957 there were 14.5 million handicraft workers and 2.4 million construction workers (Liu and Yeh 1963). According to Perkins's table (1967, p. 41), the former had a net output of 5.38 billion 1952 yuan, or 370 per head. We may assume, though this is open to question, that the construction workers were much more skilled and better equipped than the average handicraftsman. But even if we double the figure for net product per head, this only gives us a net product of 1.8 billion yuan. The use of locally produced materials, however, may be expected to be substantial. It will be assumed that the value of other industrial products used in construction equaled that of cement and timber, leaving (in 1957) 6.24 billion to represent construction wages plus use of local materials, or 2.46 times the value of timber and cement output. The outputs of these two commodities are the only indicators at present available for later years, and we will multiply their combined sum by 2.46 to obtain an estimate of net value added in construction.

TABLE 7

OUTPUT OF INDUSTRY AND CONSTRUCTION (BILLIONS OF 1952 YUAN)

Year	Output of Industry and Handicrafts	Including Net Value Added in Construction	Machinery Production (including Defense Equipment)	Gross Value of Construction	Output Not Machinery or Construction	Consumption Goods	Balance
1957	24.5	30.7	6.2	11.3	13.2	10.3	2.9
1958	32.1	40.1	11.1	14.4	14.6		
1959	40.7	50.0	15.6	16.7	17.7		
1960	40.5	48.0	18.9	13.6	15.5		
1961	25.8	31.7	9.4	10.6	11.7		
1962	26.0	32.2	7.5	11.2	13.5		
1963	29.0	36.1	9.0	12.8	14.3		
1964	32.9	40.5	10.4	13.8	16.3		
1965	37.6	46.1	14.6	15.3	16.2	10.9	5.3
1966	43.2	52.1	17.5	16.1	18.5		
1967	34.5	42.4	14.0	14.3	14.1		
1968	42.0	50.5	17.5	15.3	17.7		
1969	49.6	(59)	21.4	(16)	(21.6)		
1970	56.3	(66)	24.5	(17)	(24.5)	20.3	4.2
1971	62.0	(72)	28.9	(17.5)	(25.6)	23.4	2.2

The apportionment of output in table 7 is based on Perkins's table. Output of industry and handicrafts is from the Field-Cheng indexes (Field 1970*b*, Cheng 1973), and that of machinery, etc., from Cheng (1973).

The balance shown in table 7, after deducting consumption goods, should represent exports, stockpiling, and government purchases of goods other than capital goods or defense equipment.

Allowing for the uncertainty of all the figures, this is probably about as good an apportionment as we can get of the uses of the final output of industry and construction.

Mainland China, for its size, has in recent years done very little international trade—about as much as Taiwan. Even in an earlier period, when China was not free to pursue independent economic policies, Hou (1963) estimated that the volume of export trade was expanding at the rate of 2.4 percent per year—below the world average—between 1867 and 1932. Eckstein (1973) estimated that not until 1955 was the 1928 volume of international trade recovered. There was a rapid expansion of international trade (at that time largely with Russia) up to 1959, followed by a heavy fall consequent upon the agricultural collapse of the early 1960s.

Nonagricultural exports (largely textiles) rose from $0.5 billion in 1955 to $1.2 billion in 1959; since then, they changed little for some years but have risen again recently. The fall in exports in the early 1960s and their subsequent recovery were both due to changes in agricultural exports (see table 8).

TABLE 8

International Trade (Billions of US$)

	1955	1959	1962	1966	1972
Imports	1.31	2.06	1.15	2.14	2.77
Exports	1.43	2.23	1.52	2.21	2.95
Composition of imports (%)					
Machinery and equipment	27.0	47.0	10.5	21.0	
Agricultural products	7.5	11.4	43.3	29.0	
Agricultural products as % of exports	66.2	46.9	23.3	55.6	60.0

Sources.—Cheng (1973)—38 estimates, Eckstein (1973)—44 estimates

TABLE 9

Agricultural Imports (Millions of Tons)

	1969–70	1970–71	1971–72	1972–73	1973–74	1974–75
Grain	5.1	3.7	3.0	6.1	9.2	9.0
Soya beans					0.7	0.5
Cotton	0.07	0.10	0.15	0.42	0.45	0.3

At one time, external resources were used as far as possible for the purchase of machinery and equipment, but in the near-famine conditions of the early 1960s they had to be diverted to food imports, principally of wheat. Since 1961, grain imports have ranged between 3 and 9 million tons per year. Against this, China has maintained rice exports (except for 1961 and 1962) at the rate of about three-quarter million tons per year, but these had only been returning (until the recent sharp rise in rice prices) about $100 million per year. The agricultural products, which still constitute some 60 percent of China's exports, consist mainly of oilseeds, tea, etc. Food sold to Hong Kong (which relies principally on mainland China for fresh meat, vegetables, etc.) constitutes a substantial element in the Chinese export total.

Since 1971 China has also become an importer of cotton, and grain imports have been rising (see table 9).

According to one estimate ("PRC Economy in 1973" 1974) China's international trade increased by 84 percent in value in 1973, mainly as a result of price increases.

"Economic Indicators for China" (1974), however, makes a much lower estimate, of exports rising to $4 billion and imports to $4.5 billion.

In addition to Hollister's pioneering study, several more recent attempts have been made to express China's entire national product in terms of U.S. dollars (table 10).

Except for "Economic Indicators for China," which specifies dollars of

TABLE 10

CHINESE GROSS NATIONAL PRODUCT (BILLIONS OF US$)

Year	"Economic Indicators for China" (1974)	Klatt (1973)	World Bank Atlas	Ichimura	Tobin (1974)	U S Congress Joint Economic Committee (1968) (Factor Incomes)	Eckstein (1966)	Swamy (1973) (Billion Rupees)
1952	64	40						126 0
1957	88	55				65	40	160 2
1958	102	66						177 0
1959	96							172 6
1960	94			38 8			50	145 3
1961	79							158 6
1962	88						42	166 3
1963	93							179 5
1964	103							195 0
1965	114	70	85		80	73 3		198 6
1966	122							200 8
1967	119			50				
1968	118		78					
1969	130							
1970	147	90			112			
1971	159							
1972	161							
1973	172							

1972 purchasing power, and Eckstein (1966), who indicates dollars of 1961 purchasing power, it appears that each author was thinking in terms of the value of the U S dollar at the time he was writing, not its (higher) value in any past year to which he referred

Tobin (1974) begins by observing that the current exchange rate of $0 45 to the yuan "if anything understates the dollar value of the yuan in buying a Chinese market basket " He then applies this rate to value China's "material product" at $90 billion, to which he adds $12 billion for services (including government), occupying an estimated 10 percent of the labor force, and a further $10 billion for imputed rents of dwellings, free hospitalization, etc

Tobin pointed out that 80 percent of the population was still rural, but estimated that agriculture contributed only 40 percent of the national product (He did not refer to industrial and constructional activities also carried on by rural communes, perhaps, however, they do not add up to too much) Klatt (1970) makes agriculture contribute 45 percent of the national product in 1952, 40 percent in 1957, 31 percent in 1965, and only 29 percent in 1970. Eckstein (1973) gives similar figures. he estimates 75 percent of the labor force in agriculture in 1970

However, for international comparisons these figures are much too low, principally because of the low valuation which they put on agricultural production. It is true that the physical productivity of Chinese agriculture is low. But we must give it credit for the fact that, since most of the food is produced and consumed in the same village, Chinese agriculture does not incur the costly processing, transport, packaging, and distribution charges which arise in the advanced countries. For fair comparison, therefore, we should value Chinese food supply at U.S. *retail* prices, if we wish to revalue the national product in dollars. (This has an additional effect in raising the Chinese valuation, in that prices of vegetables and animal products, relative to staple grains, are considerably higher in the United States than in China.)

Using the above method, and all other information available of the purchasing power of the yuan relative to dollars over different types of goods and services, I obtained valuations (much higher) of the Chinese national product in 1933 and in 1950–59 (Clark 1965). The unit was the purchasing power of the U.S. dollar in 1950 (this year was chosen because it was the base year for the well-known Gilbert-Kravis international comparisons of the purchasing power of different currencies). If we wish to reexpress these results in 1974 dollars, we must multiply by approximately 2.1.

However, the estimates published in 1965 need revision in three respects. (1) Vegetable supplies (following Hollister 1958) were estimated at only 15 million tons for the 1950s. Klatt (1967) estimates fruit and vegetable supplies in 1966–67 at 50 million tons, or approximately 70 kilograms per head. (At the yield of 37 tons per hectare mentioned above for at least the more productive plots, only 1.3 percent of the cultivated land would suffice to produce 50 million tons. Indeed, if [as now appears to be the case] private plots amount to over 5 percent of the cultivated land, and if they are largely used for vegetables, even the estimate of 50 million tons may be too low.) Ishikawa (1965) reports average consumption by Shanghai workers' families (whose real incomes, however, were above the national average) at 97 kilograms per head per year. The calculations were therefore adjusted to raise estimated vegetable consumption in the 1950s (and also in the 1930s) to 45 million tons per year. (2) Gross investment (i.e., before depreciation) is substituted for net. Gross investment in machinery production is also revised upward in view of Cheng's (1973) estimates, though Hollister's conversion factor of 0.18 from 1952 yuan to 1950 dollars is retained. (3) The purchasing power of the yuan for consumption other than food or housing was previously given (Clark 1965) at $1,620 of 1950 purchasing power for 1955 (base calculation) adjusted for other years between 1950 and 1958 by the general retail price index, with a 1953–58 average of 0.619. The evidence on which this was based was limited, and it is now estimated that it should be raised by a factor of 1.52, in the light of table 11 (Japanese price comparisons quoted by Swamy [1973]).

TABLE 11

PURCHASING POWER (CONSUMPTION NOT FOOD OR HOUSING) OF THE YUAN, 1953–58

	Yen 1953–58 / Yuan 1953–58	Yen 1960* / 1960$	Yen 1960† / 1950 $	Yen 1953–58‡ / 1950 $	Yuan 1953–58 / 1950$	Weight§
Clothing	59	182	206	209	0 282	0 427
Fuel	556	285	344	320	1 74	0 277
Other goods	160‖	167	194	177	1 09#	0 296
Services				85†		
Weighted mean					0 928	

* *Japan Economic Research Council Bulletin*, vol 12 (1963) and Kumano (1961) (for consumer durables)
† Previous column adjusted by U S national product deflators
‡ Previous column adjusted by Japanese national product deflators
§ Liu and Yeh (1963, table 10) (for 1952)
‖ "Daily necessaries," consumer durables and "cultural" combined with weights 1 1 2
# Previous column combined with the following weights 2 for goods, 1 for services

TABLE 12

FOOD SUPPLIES

	Millions of Tons, 1966 67	Billions of Dollars of 1950 Purchasing Power, 1966-67	1970
Rice*	77 5	9 6	20 5
Wheat*	24 5	3 9	
Other cereals*	33 0	4 32	
Potatoes	48 0	1 15	
Soya beans	10 5	3 26	33 7
Sugar	1 5	0 34	
Fruit	5†	3 05	
Vegetables	45†	9 7	
Meat and poultry	7	5 25	
Fish and eggs	6	4 5	
Oils and fats	3 5	2 8	
Total		47 87	54 2

* Unmilled weight
† Arbitrary apportionment of Klatt's 50-million-ton total

Klatt (1967) presents a food balance table for 1966-67, taking into account imports, exports, seed, waste, fodder use, industrial requirements, etc , from which a valuation can be made in 1950 dollars for 1966–67, and a further very approximate valuation for 1970 (see table 12)

Available indications show 1965 supplies as about the same as in 1966–67 A valuation has also been made for rural fuel, in which the villages supply most of their minimum requirements without purchasing from industry Buck (1964) estimated that the average Chinese farm family, in addition to supplying 83 percent of its food and the whole of its housing requirements, also provided 89 percent of its fuel, 18 percent of its clothing, and 5 percent of its furniture.

The purchasing power of the yuan is particularly high in residential construction, where the productivity of labor does not differ from that of the advanced countries as much as it does in manufacture and agriculture. In this sector the purchasing power of the yuan in the 1950s was estimated at $1.5 of 1950 purchasing power (Clark 1965).

We can now revise the estimates published in 1965 (Clark 1965) (to convert investment from net to gross, to make an increased allowance for vegetable production, and to include a revised valuation of the purchasing power of the yuan over consumption other than food and housing) and bring them approximately up to date (see table 13).

In the first place, food supply per head, even in the best years of the 1950s, had only just recovered to the 1933 level, and since then has been below it, in 1960–63 far below it.

National product per head as a whole is shown in figure 1. Some of the years of rapid growth represented nothing more than recoveries from the economic disaster of "The Year of the Great Leap Forward" and the lesser setback of the Cultural Revolution, and cannot therefore be taken as indicative of the general long-period rate of growth. This may be gauged by comparing product per head in 1970–71 with 1951–52, or with 1933—namely, 1.9 and 2.0 percent, respectively, per head per year (compounded).

These rates of growth are much below the rates usually claimed, and below the average rate of growth of most other developing countries.

Finally, we may check these results against such information as is available on wages and prices.

There has been some discussion as to whether there has been any price rise in China. The prices analyzed below, however, indicate (for goods other than food or rents) a rise of only about 4 percent per year.

A black market did indeed develop in the 1950s (*China News Analysis* 1956), when the official price of rice in South China was 0.26 yuan per kilogram, but black market rice on one occasion was selling at 0.9 yuan per kilogram. We are also given a rare example of the Chinese sense of humour. "The man in the shop said 'This shop is Chairman Mao's, therefore we have no fish, but go to Liu Shao-Ch'i's, he may have some.' "

From recent information (from a private source) on retail prices in China, we can value Chinese private consumption expenditure on nonfood goods and services in terms of current (or 1950) U.S. prices. The bases from which the U.S. general consumer price index is calculated are not published in a form convenient for comparison. Base U.S. prices (for 1971) are therefore taken from detailed tables of "Prices Paid by Farmers" (U.S. Department of Agriculture 1972). As the Chinese specifications did not directly fit the U.S. base data, a "bridge" was obtained, using current

TABLE 13

CHINESE GROSS NATIONAL PRODUCT REVALUED IN U S DOLLARS OF 1950 PURCHASING POWER (BILLIONS)

	1933	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971
Food supply (urban and rural)[*]	45.8	42.5	43.1	43.8	45.2	46.4	46.7	48.0	47.3	48.7	50.6	41.5	39.8	43.0	44.6	48.0	47.9	46.8	49.8	49.8	49.8	54.2	53.8
Fuel (rural)	1.2	1.2	1.2	1.2	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3
Rental value of housing (urban and rural)[†]	7.3	7.3	7.3	7.2	7.3	7.4	7.6	7.7	7.9	8.1	8.1	8.2	8.2	8.2	8.3	8.3	8.4	8.5	8.6	8.7	8.8	9.0	9.1
Gross investment																							
Construction[‡]		5.2	5.4	6.9	8.1	9.2	10.0	14.5	15.2	19.3	22.4	18.3	14.2	15.0	17.2	18.5	20.5	21.6	19.2	20.6	21.5	22.8	23.5
Machinery[§]	12.9																						
Production[‡]		0.3	0.3	0.4	0.4	0.6	0.9	1.4	1.1	2.0	2.8	3.4	1.7	1.3	1.6	1.9	2.6	3.2	2.5	3.2	3.9	4.4	4.6
Imports		0.2	0.3	0.3	0.3	0.4	0.4	0.5	0.5	0.6	1.0	0.9	0.3	0.1	0.2	0.2	0.3	0.4	0.3	0.2	0.2	0.3	0.4
Inventory change and net foreign investment		0.6	1.5	1.7	0.9	1.1	−0.1	−0.4	2.1	3.2	2.0	0	−2.5	−2.5	1.0	1.0	1.0	1.5	1.5	1.5	1.5	1.5	1.5
Other private consumption[‡]		9.3	11.8	15.8	17.2	19.3	18.6	20.5	19.7	21.9	26.4	23.3	17.5	20.7	21.4	24.5	24.4	27.6	21.1	26.5	32.4	36.7	38.4
Government services[§]		8.3	12.7	11.5	13.3	13.0	13.0	12.5	11.4	12.3	14.4	16.5	12.5	12.5	13.5	14.7	15.7	18.5	12.0	15.0	21.0	25.0	26.0
Gross national product	67.2	74.9	83.6	88.4	94.0	98.7	98.4	106.0	106.5	117.4	129.0	113.4	93.0	99.6	109.1	118.4	122.1	129.4	115.5	126.8	140.4	155.2	157.6
Population (millions)	590	556	565	574	583	593	603	613	624	635	645	650	650	656	664	672	680	688	697	706	715	724	733
Product per head																							
Food	77.5	76.3	76.4	76.4	77.6	78.3	77.4	78.4	75.8	76.8	78.5	63.8	61.2	65.5	67.2	71.4	70.4	68.0	71.5	70.5	69.7	74.9	73.4
Other	36.4	58.2	71.6	77.7	83.8	88.3	85.8	94.6	94.9	108.2	121.5	110.5	81.8	86.2	97.2	104.6	109.2	120.2	94.4	109.0	126.7	139.6	141.7
Total	113.9	134.5	148.0	154.1	161.4	166.6	163.2	173.0	170.7	185.0	200.0	174.3	143.0	151.7	164.4	176.0	179.6	188.2	165.9	179.5	196.4	214.5	215.1

[*] To 1962 from Clark 1965, 1965 and 1970 calculated in text Other years interpolated on grain output

[†] No information is available after 1959, but strong measures were taken to check the growth of urban population, because of the housing shortage, among other reasons.

[‡] Old *China Quarterly* data (revised valuation) extrapolated 1957–71 from table 7

[§] Including defense equipment Figures for 1964 and 1965 from Wenmohs (1967), others interpolated on private consumption

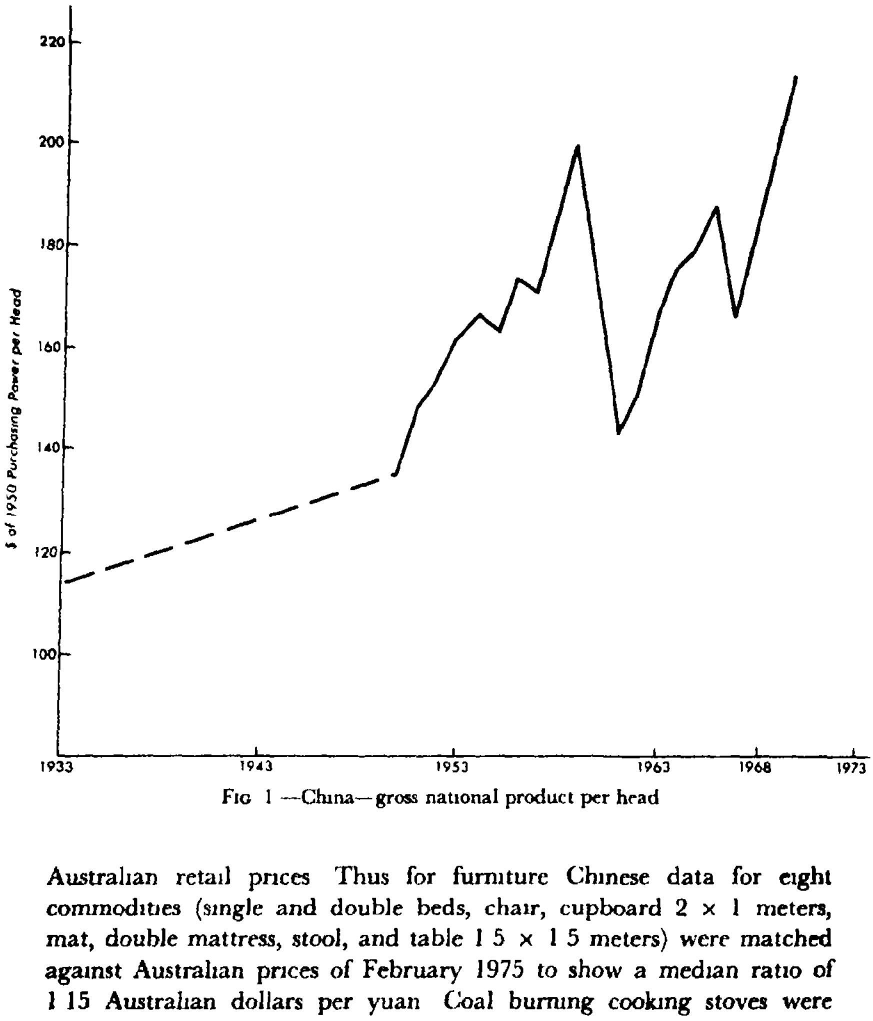


FIG 1 —China—gross national product per head

Australian retail prices Thus for furniture Chinese data for eight commodities (single and double beds, chair, cupboard 2 × 1 meters, mat, double mattress, stool, and table 1 5 × 1 5 meters) were matched against Australian prices of February 1975 to show a median ratio of 1 15 Australian dollars per yuan Coal burning cooking stoves were specified in China but are no longer on the market in Australia Then an Australian-U S comparison was made on different (higher-quality) specifications of innerspring mattresses and box spring sets, bedroom sets, living room sets, dining room sets, and upholstered lounge chairs, showing a median ratio of 1 23 Australian dollars (1975) per U S. dollar (1971), indicating the purchasing power of the yuan in this sector at US$0.93 of 1971 purchasing power (table 14)

Medians rather than averages are used throughout, as they are less affected by extreme values and specification errors. For services, direct U.S. price information was taken from a United Nations document (United Nations 1971) referring to New York in November 1969, and

TABLE 14

INTERNATIONAL PRICE COMPARISONS

	No of Commodities for which Chinese Prices Available	Median. Australian $/ Yuan	Median. Australian $/ US $	US $/ Yuan	Weights*
Goods					
Clothing and shoes	21	0 99	1 64	0 60	0 427
Fuel	5	1 38	1 80	0 77	0 277
Furniture	8	1 15	1 23†	0 93	0 018
Minor household equipment	29	0 60	0 90	0 32	0 038
Tobacco	1	1 59	1 54	1 03	0 064
Other recreation	15	0 59	1 43	0 41	0 042
Household textiles	7	0 45	1 52	0 30	0 017
Toilet goods	5	0 40	1 30	0 31	0 017
Total goods (weighted mean)				0 657	
Services	11			7 7	0 100
General weighted mean				1 362	

* The two major items, clothing and fuel, are from Liu and Yeh (1963 Table 10) (for 1952) For services, an arbitrary estimate was made The apportionment of the remaining items was estimated from Indian evidence

† For mechanical household equipment (not included in the calculation,) the ratio was 1 33

from some personal information (to convert to 1971 values, the 1969 data were raised by approximately 10 percent) Postal and telephone services are comparatively costly in China, recreational services and hairdressing (as was to be expected) extremely cheap There have been complaints, however, of hair cuts (normal charge 0 2 yuan) costing 1–2 yuan in a remote area in Sin Kiang ("Interview" 1961) Medical treatment is not free but is charged for at the rate of 1 5–2 yuan (ibid) Nor is schooling free, though it is clearly subsidized—Middle School students were charged 11 yuan per term (" 'Who Is Mao Tse-tung?' " 1962) From other sources, there are indications that primary schools also charge fees, though at a much lower level—about 10 yuan per year inclusive of charges for books and stationery Rents for government-provided dwellings ("Workers' Housing" 1956) averaged 5–10 percent of wages In the 1950s there was still some private renting, rents averaging about 20 percent of wages For food Middle School boarders are charged 10½ yuan per month, and in factory canteens ("Writer Returns" 1971) 15 yuan per month An earlier report ("The Janus-Face of Labour Conditions" 1956) also indicates textile workers without dependents and eating in factory canteens spending 30–33 percent of their wages on food, that is, 15–16 yuan per month in terms of wages of the 1950s.

If we estimate that 50 percent of the urban population is working (i.e., an average of one dependent per worker), food for one adult and one child, on the above scales, would come to 20 yuan per month.

This conflicts with the previous (Clark 1965) general estimate of urban wage-earning families spending 65 percent of their incomes on food (50 yuan in terms of average earnings in the 1950s) Factory and school canteen food may, however, be below the average standards of consumption of the urban population

It is interesting to observe the high purchasing power of the yuan over the most labor-intensive commodity (furniture) and, to a less extent, over clothing, as well as its extremely high purchasing power over services by U S standards Its high purchasing power over fuel, and still more over tobacco, is the result of administrative decisions—in the latter case, not to tax the poor man's principal recreation as severely as do the advanced countries

We now need to relate these purchasing powers expressed in 1971 U S dollars to those in 1950 dollars previously used (see table 15) This means that the purchasing power (excluding food and housing) of the present-day yuan, which was 0 657 for goods and 7 7 for services in terms of 1971 U S dollars, was 0 466 for goods and 3 5 for services in terms of 1950 U S dollars, or 0 769 in all, on the weights used above

For 1953–58 (see above) its purchasing power was estimated at 0 928, that is, a price rise of about 1 percent per year

We now consider the available information on earnings

The belief that earnings in China show a high degree of equality is mistaken In the 1950s (*People's Daily* 1956) average earnings of all wage and salary earners were 50 8 yuan per month, with the following range primary school teachers, 37 yuan, state farm workers, 44 8, Shanghai bank officers, 68 6, Shanghai municipal workers, 70 2, technical personnel, 88, and senior engineers, 216

Regular building workers received 47 5 yuan per month, but casual workers in 1951–52 (excavating an average of 2 3 cubic meters per day) received only 0 8 yuan, equivalent to only 4 6 kilograms of wheat per day—not much above a subsistence wage (Chao 1965, Yeh n d)

The Indian Agricultural Delegation (Government of India 1956) of 1956 reported professors earning 250–400 yuan, craftsmen, 100, laborers and minor clerical workers, 35—not so different from the Indian scale of earnings Working hours were reported at 10 per day, including "education" (8 hours' work followed by 2 hours' study, according to a party resolution of December 1958) for 306 days per year (The end of 1958 was, however, a period of extreme political tension, and this rule was probably subsequently relaxed) In the busy harvest season (which, however, does not last very long) rural workers worked a 12-hour day.

After "The Year of the Great Leap Forward" it appears that wages were reduced. A refugee coolie reported (*Current Scene* 1962). "At first I made 40 yuan/month but my wages were cut until I was only earning 30 yuan/month. I could carry 80 kilograms [!] Now I can only carry 55–60."

TABLE 15

U.S. PERSONAL CONSUMPTION

	1950			1971		
	Current (Billions of $)	At 1958 Prices (Billions of $)	Price Index	Current (Billions of $)	At 1958 Prices (Billions of $)	Price Index
Food and beverages	53.9	63.2	0.854	136.4	103.4	1.319
Housing	21.3	26.8	0.795	99.2	75.6	1.311
Other goods*	78.6	90.2	0.871	264.6	215.2	1.228
Services†	37.2	50.3	0.740	164.7	101.2	1.627
Total	191.0	230.5	0.828	664.9	495.4	1.343

* By difference from total

† Excluding housing and certain fuels recorded as "services."

In 1970 it was reported ("China's Economy in 1969" 1970) that average wages were still 50 yuan per month, but that there had been a great equalization, with maximum earnings now 100 yuan per month, in place of former figures of 150–250. This, however, conflicts with other information, showing average wages in Canton at 60–70 yuan per month ("Writer Returns" 1971) and Tobin's wider observation (1974). In a textile mill he found an average of 60 and a minimum of 35, with the highest-paid manual worker receiving 100 and engineering and technical staff receiving 130–140. In an automobile workshop the lowest paid received 40, the most experienced 200, in a department store the minimum was 36, the average 63, and the maximum 80. In the university the minimum was 50 but the maximum over 200. The average for all hospital staff was 67, including doctors averaging 85 but surgeons 200–300. However, newly qualified doctors were said to receive only 42–48 (Sieh 1964).

We have little information about the present cash incomes of the farm population. However, Pi Chao Chen (who succeeded in getting some information on population out of Chou En Lai) also was able to visit six agricultural "brigades" in 1972 (total populations ranging from 670 to 7,490 persons per family ranging from 4.8 to 6, median 5.2). One was a specialized tea-growing enterprise, with income all in cash (210 yuan per person per year), out of which they had to buy grain at the urban price of 0.28 yuan per kilogram. This income however seemed to be above the average, valuing income paid in grain at the above price, three other brigades had average per head incomes equivalent to 167, 169, and 183—a general average of 182.

Judging, as far as possible, from national averages of agricultural productivity and prices per head, it does not appear that the visitors were taken to see better-than-average brigades to any great extent. In a village in Kweichow, said to have been better off than the provincial average (but Kweichow is a relatively poor province), valuing 350 kilograms of *unmilled* rice at the urban price of 0 18 yuan per kilogram, the average family income was again placed at 183, with the less successful earning 40–50 yuan less ("Agriculture, 1968" 1968)

Tobin (1974) estimated rural incomes (including incomes in kind) at half urban levels—presumably he meant per head An average rural income of 180 yuan would thus be compatible with an average urban wage of 720 yuan and 50 percent of the population in employment (in most industrial countries the ratio is now about 40 percent, and in Chinese cities, even if there are more children, there are almost certainly more married women working and far fewer pensioners)

Incomes per head are therefore estimated at 360 yuan urban and 180 rural, or 225 in all (with 75 percent of the population rural) The National Product table shows 1970-71 consumption, other than food, rent, and rural fuel, at $37 5 billion of 1950 purchasing power, or 48 8 billion current yuan (on the yuan's purchasing power of 0 769 calculated above), that is, 67 yuan per head This is compatible with an estimate that outlays for food, rent, and savings amount to 70 percent of the average (rural plus urban) of personal incomes

References

"Agriculture, 1968 " *China News Analysis* (October 18, 1968), pp 1–7

British Broadcasting Corporation *Summary of World Broadcasts*, pt 3 (Far East), no 4077 (August 26, 1972)

Brown, Harrison, et al *Population Perspective, 1972* San Francisco Freeman, Cooper, 1973

Buck, John Lossing *Land Utilisation in China* New York Paragon, 1964

——— "Reliability of Communist China's Data on Food Grain Production." *Current Scene* (March 1, 1965), pp 1–11

Chao, Kang "Indices of Industrial Output in Communist China " *Rev Econ and Statis* 45 (August 1963) 286–93

——— "Growth of the Construction Industry in Communist China " *China Q.*, no 22 (April–June 1965), pp 125–42

———. *Agricultural Production in Communist China 1949–1965* Madison Univ. Wisconsin Press, 1970

Chen Pi-chao "China's Population Program at Grass Roots Level " *Studies in Family Planning* 4 (August 1973) 219–27

Cheng Chu-yuan. "China's Machine Building Industry." *Current Scene* (July 1973), pp. 1–11.

China News Analysis (July 19, 1956)

"China's Economy in 1969 Policy, Agriculture, Industry, Foreign Trade." *Current Scene* (June 1, 1970), pp 1–17

Clark, Colin "Economic Growth in Communist China." *China Q.*, no 21 (January–March 1965), pp 148–67.

———. "La population de la Chine depuis 1915 " *Population* (Paris) (November–December 1966), pp 1191–99.

Clark, Colin, and Haswell, Margaret R. *The Economics of Subsistence Agriculture*. 4th ed. London. Macmillan, 1964.

Communist China. Pt. 2. Hong Kong: Union Research.

Eckstein, Alexander. *Communist China's Economic Growth and Foreign Trade · Implications for U.S. Policy.* New York: McGraw-Hill, 1966.

——— "Economic Growth and Change in China · A Twenty-Year Perspective." *China Q* , no 54 (April-June 1973), pp. 211–41

"Economic Indicators for China " *U S China Business Review* 1 (March–April 1974): 33.

Field, Robert Michael "Comment Chinese Provincial Population Data." *China Q* , no 44 (October-December 1970), pp. 195–202 (*a*)

——— "Industrial Production in Communist China 1957–1968 " *China Q* , no. 42 (April-June 1970), pp 46–64 (*b*)

Government of India Report of Agricultural Delegation to China 1956

Hollister, William Wallace *China's Gross National Product and Social Accounts, 1950–57* Glencoe, Ill Free Press, 1958

Hou Chi-ming "Economic Dualism The Case of China, 1840–1937 " *J Econ Hist* 23 (September 1963) 277–97

"Interview 'The Cadres Nicknamed Me "Erh Liu Tzu"' . " *Current Scene* (November 18, 1961), pp 1–18

Ishikawa, A *National Income and Capital Formation in Mainland China* Tokyo. Inst Asian Econ Affairs, 1965

"The Janus-Face of Labour Conditions " *China News Analysis* (April 20, 1956), pp 2–7 *Japan Economic Research Council Bulletin,* vol 12 (1963)

Klatt, Werner "China's Third Five-Year Plan " *China Q* , no 25 (January-March 1966), pp 171–75

——— *Analyse et prévision* (Paris) (1967), p 841

——— "A Review of China's Economy in 1970 " *China Q* (July-September 1970), pp 100–20

——— "A Hungarian View of China's Economy " *China Q* , no 54 (April-June 1973), pp 341–44

Kumano, Hideaki "American Dollar and Japanese Yen Purchasing Power Equivalents " *J Econ Behavior* (Japan) 1 (October 1961) 113–26

Liu Ta-chung, and Yeh Kung-chia "Preliminary Estimate of the National Income of the Chinese Mainland 1952–59 " *A E R* 51 (May 1961) 489–98

——— *Economy of the Chinese Mainland* Santa Monica, Calif RAND, 1963

Mao Tse-tung *Socialist Upsurge* Peking Foreign Languages Publishing House, n.d.

Nihon keizai shimbun (English ed) (February 4, 1975)

Orleans, Leo A "Propheteering—the Population of Communist China " *Current Scene* (December 15, 1969), pp 13–19

——— *Every Fifth Child The Population of China* London Methuen, 1972

——— "China's Population—Some Confirmations and Estimates " *Current Scene* (March 1974), pp 13–15

People's Daily, Peking (July 17, 1956)

Perkins, Dwight H "Economic Growth in China and the Cultural Revolution (1960 April 1967) " *China Q* , no 30 (April-June 1967), pp 33–48

——— "Mao Tse-tung's Goals and China's Economic Performance." *Current Scene* (January 7, 1971), pp 1–15.

"PRC Economy in 1973." *Current Scene* (March 1974), pp 1–11.

"Private Plots and Three Level Ownership " *Current Scene* (May 7, 1971), pp 17–18

Rawski, Thomas G "Chinese Industrial Production 1952–71 " *Rev. Econ. and Statis* 55 (May 1973) · 169–81. (*a*)

——— "Recent Trends in the Chinese Economy." *China Q.*, no. 53 (January-March 1973), pp. 1–33. (*b*)

Sieh, M "Medicine in China." *Current Scene* (October 15, 1964), pp. 1–12.

Snow, Edgar "Interview with Mao Tse-tung " *New Republic* (February 27, 1965), pp 17-23

Spengler, Joseph J "Review of *Studies in the Population of China, 1368-1953,* by Ping-ti Ho." *Comparative Studies in Society and History* 5 (1962-63). 112-14

Stanford Research Institute *Economic Potential of Communist China*

State Statistical Bureau *Ten Great Years.* Peking Foreign Languages Publishing House, 1960.

Swamy, Subramanian "Economic Growth in China and India, 1952-1970: A Comparative Appraisal." *Econ Development and Cultural Change,* vol 21, no. 4, pt. 2 (July 1973).

Tobin, James "The Economy of China—a Tourist's View " *Current Scene* (May 1974), pp 1-12

United Nations *Retail Price Comparisons for International Salary Determination* Ser. M, no 14, rev 1 New York Dept. Econ and Soc Affairs Statis Office, United Nations, 1971

U S Congress Joint Economic Committee *An Economic Profile of Mainland China* New York Praeger, 1968

U S Department of Agriculture *Agricultural Statistics, 1972* Washington Government Printing Office, 1972

Walker, Kenneth Richard *Planning in Chinese Agriculture Socialisation and the Private Sector, 1956-1962* London Cass, 1965

Wenmohs, John R "Agriculture in Mainland China—1967 Cultural Revolution versus Favorable Weather " *Current Scene* (December 15, 1967), pp 1-12

"'Who Is Mao Tse-tung?' Four Answers " *Current Scene* (January 5, 1962), pp 1-10

"Workers' Housing " *China News Analysis* (December 14, 1956), pp 2-3

"Writer Returns Finds Old Raucous Canton Has Turned Sedate " *Chicago Tribune* (May 22, 1971)

Yeh Kung-chia RAND Corp Paper P. 3251 Santa Monica, Calif RAND, n d.

A Stochastic Model of Applied Research

Robert E. Evenson

Agricultural Development Council, College, Laguna, The Philippines

Yoav Kislev

The Hebrew University, Rehovot

A mathematical model of applied research is formulated. It views applied research as a search in a given distribution, basic research shifts the distribution searched. The productivity of applied research effort is a function of the gap between technology in practice and basic knowledge. With constant basic and applied research a (stochastic) steady state emerges in which technological change is determined by the rate of progress of basic knowledge, and the technological gap by the level of applied research.

Much of research work is experimentation, and often a technological development project consists of the testing of a collection of technologies (methods, formulas, timing, varieties of crops) to find the best one. At more basic levels of research, the scope of technologies available for testing is increased. Better theoretical understanding results in superior technologies.

This experimentation process is formulated mathematically in this

---

This study was supported by a grant from the United States–Israel Binational Science Foundation. Previous support by the International Bank of Reconstruction and Development and by the Economic Growth Center of Yale University is also acknowledged. We have benefited from presentation and discussions of early drafts of this paper in seminars at Rochester University and the Airlie House Conference on Resource Allocation and Productivity in International Agricultural Research. We are indebted to Meir Kohn, Hanna Lifson, Richard R. Nelson, and the *Journal* readers for comments and important suggestions.

[*Journal of Political Economy*, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved.

paper The approach is illustrated mostly in terms of agricultural research but is of a wider applicability

The approach was first suggested by the study of the development of sugarcane varieties (Evenson and Kislev 1975, chap 3) It is inspired by Stigler's work on "The Economics of Information" (Stigler 1961) and is similar in some aspects to Nelson's treatment of R & D (Nelson 1961)[1]

Four stages were identified in the sugarcane breeding history First, sexual reproduction of the cane plant was not known, and improvements were very slow, based on the occasional and rare cases of "natural" sexual reproduction The second stage was the discovery of the conditions necessary to induce flowering and sexual reproduction Seedlings were drawn from random crossings, and plants with superior potential were selected and propagated vegetatively The third stage introduced purposeful crossing, directed to augment desired traits The fourth stage was marked by even more specific selection programs directed to produce varieties suited to local soil, climate, and technological conditions

Selection can be described in statistical terms as a process of random drawings from a distribution The overall distribution in sugarcane is that of all possible genetic combinations and mutations of the species

In the first stage of the cane breeding, the distribution sampled was that of the vegetatively reproduced, and therefore identical, plants with only occasional crossing Thus, the drawing was limited to very few distinct observations The population sampled was heavily concentrated around the current yield level

In the second stage, drawing was from the distribution of all possible crossings Compared with the first, this stage was marked by an enormous increase in the *sample variance* The third and fourth stages were characterized by the development of techniques to affect the distributions searched, to limit it to the part of the population with the stronger desired characteristics, and thus to shift its mean

Diminishing returns set in as search continues within the same distribution, and productivity of research increases when the search shifts to a new population This was exemplified by the history of sugarcane varietal development in Barbados, British West Indies Work on Stage II varieties was carried out from 1880 to 1939 The station released 10 important commercial varieties from 1902 to 1912, one from 1914 to 1928, and one important and three minor varieties from 1928 to 1939

---

[1] For recent important contributions to the theory of economic search see Kohn and Shavell (1974) and Rothschild (1974) A less known early contribution is Karlin (1962) (pointed out by an anonymous *Journal* reader) Binswanger (1974) utilized a similar framework (based on an earlier draft of the present paper) in constructing his model of the micro-theory of induced innovation.

Work on Stage III varieties started in 1929. Under this project, 14 commercially important varieties were released between 1929 and 1939.

The diminishing returns to search in the same population are also demonstrated by the increasingly larger numbers of cane seedlings tested per successful "find" of a commercial variety. During the period 1929–39, the ratio of successful varieties to seedlings tested was 1:1,800 for the first five and 1:2,700 for the next nine Stage III varieties. During the same period and in the same station, this ratio was 1:13,000 in the older Stage II variety improvement program.

Technological research in other fields can be described in similar terms (e.g., the development and selection of chemical compounds). In our model, applied research is seen as a search within a distribution of a random variable. Basic research or learning shifts the mean of the distribution or discovers new distributions to search. An intermediate position is occupied by work aimed at increasing the variance of the samples. The model formulated is an economic one; it is assumed that the objective function of the system is to maximize the expected present value of future income, inclusive of research cost. The system modeled can be either an economy with a research sector, or limited section of the economy with a research team working to improve technology within the section.

The main innovation of the study is in the treatment of applied research and testing, emphasis in the following discussion being given to this aspect of research work. The framework of the analysis is initially to set up a pure-search model and later to introduce the possibility of more basic research into the model.

Applied Research (Testing)

The scientist (the scientific team) is assumed, in this section, to be presented with a given distribution of outcomes whose parameters he cannot directly affect; his work is strictly testing, no basic research is done. To be concrete, imagine a research project aimed at increasing the yield of a crop. To simplify, assume that net income is in direct proportion to yield. Work on the project is composed of a succession of experiments. In control theoretic language, the state of the system is the yield at any point in time, and the results of the experiments are the transition equations—changing the yield level. The control variable is the extent of experimentation at any stage or time period. Again for simplicity, assume that the only control variable is the number of trials in an experiment—the number of drawings from a random distribution.

Since the transition equation is a random process, the state variable is random, but other sources of randomness and uncertainty, such as weather effects, are disregarded.

At any period t an experiment composed of n_t trials is conducted. The following variables are defined as follows

Y_t	yield, technology level, net income in time t ($t = 0, 1, 2, \quad$), the *state* variable,
n	number of trials, $n = 0, 1, 2, \quad$, the *control* variable,
$c(n)$	cost of experimentation, with $c(0) = 0$, and assume that $c(n)$ is increasing with n at an increasing rate,
x_i	yield in trial i, ($i = 1, 2, \quad , n$),
$f(x)$	probability density function of x,
$F(x)$	cumulative distribution of x,
z	the largest value in a sample of the random variate x,
$\alpha = \dfrac{1}{1 + r}$	the discount factor with r the rate of interest,
V	the objective function,
E	expectation operator,
$\Delta y = y_t - y_{t-1}$	yield increment,
$E\{x\}$	mean of x,
Var (x)	variance of x,
D	first difference operator, e g , $D_n c \equiv c(n) - c(n - 1)$

The search process is a sequence of *experiments*, each composed of n_t *trials* A single trial can be a test of a technique—one variety of a crop, a certain dose of fertilizers, one planting date Because of the variability in experimental conditions, a trial is usually carried out in a number of replicas This variability is, however, disregarded here, and it is assumed that a trial has a single outcome—an observation from the distribution of yields, in our example

Each trial results in an observation—one drawing from a random population The outcome of the experiment is the best observation in the sample The statistical process of choosing the best outcome from a set of random drawings is treated under the heading of the theory of extreme values (Gumbel 1958, Epstein 1960) in the general subject of order statistics

Utilizing the symbols introduced earlier, x_i is the yield in trial i and $z = x_j, x_j \geq x_i$ ($i = 1, 2, \quad , n$)

The cumulative distribution of z is

$$H_n(z) = \Pr(\text{all } x_i \leq z) = F^n(z), \tag{1}$$

and the density function (if existing) is

$$h_n(z) = nF^{n-1}(z)f(z). \tag{2}$$

The analysis will be illustrated in terms of the exponential distribution:

$$f(x) = \lambda e^{-\lambda(x-\theta)}, \quad \theta \leq x; \tag{3}$$

$$F(x) = 1 - e^{-\lambda(x-\theta)}, \tag{4}$$

$$E(x) = \theta + \frac{1}{\lambda}, \tag{5}$$

$$\text{Var}\,(x) = \frac{1}{\lambda^2}. \tag{6}$$

The cumulative distribution of the largest values is, employing (1),

$$H_n(z) = [1 - e^{-\lambda(z-\theta)}]^n, \tag{1'}$$

and the probability density function is

$$h_n(z) = \lambda n[1 - e^{-\lambda(z-\theta)}]^{n-1} e^{-\lambda(z-\theta)} \tag{2'}$$

See figure 1 where $\lambda = 1, \theta = 0; n = 1, 2, \quad, 5$

The expected value and the variance of z are (Gumbel 1958)

$$E_n(z) = \theta + \frac{1}{\lambda}\sum_{i=1}^{n}\frac{1}{i}, \tag{7}$$

$$\text{Var}_n\,(z) = \frac{1}{\lambda^2}\sum_{i=1}^{n}\frac{1}{i^2} \tag{8}$$

In each time period, an experiment consisting of n trials (n drawings from a random distribution) is conducted The result of each trial is an observation, x, the yield level associated with the tested technique. If the outcome of the experiment is higher than y, the yield under the current technique, the new technique is put to use and y increases If not, y does not change The search is then repeated, within the same distribution, perhaps with a different n

Formally,

$$\Delta y = z - y \quad \text{if } y < z \tag{9}$$

$$\Delta y = 0 \qquad \text{otherwise}$$

The expected value of the technology increment is[2]

$$E_n(\Delta y) = \int_y^\infty (z - y)h_n(z)\,dz = \int_y^\infty [1 - F^n(z)]\,dz \tag{10}$$

---

[2] Strictly speaking, the last integral in (10) exists for F such that $F^n > 1 - z^{-2}$ for large values of z, which holds for the exponential distribution For a more general treatment of this case, see De Groot (1970, p. 246). We are indebted to a *Journal* reader for this reference

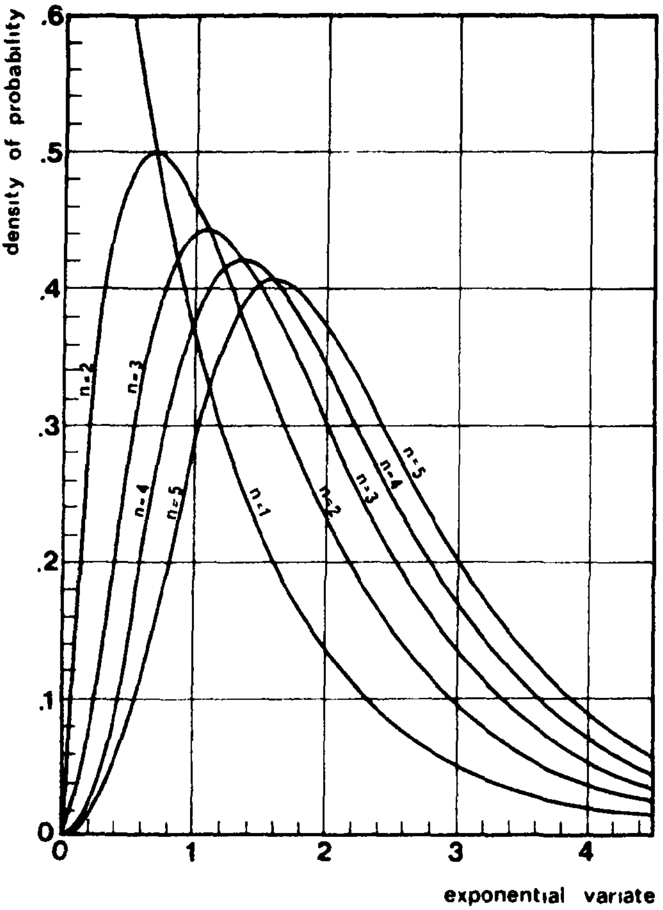


FIG 1 —The function $h_n(z) = n(1 - e^{-z})^{n-1}e^{-z}$ Source· Gumbel 1958

As expected, the contribution of an additional trial is positive and diminishing

$$E_n(\Delta y) - E_{n-1}(\Delta y) = \int_y^{\infty} F^{n-1}(z)[1 - F(z)]\,dz \qquad (11)$$

Since $0 < F(z) < 1$ for $y < z < \infty$, the difference $E_n(\Delta y) - E_{n-1}(\Delta y)$ is positive and is a decreasing function of n

For the exponential distribution the expected value of Δy is

$$E_n(\Delta y) = \sum_{i=1}^{n} \frac{1 - [1 - e^{-\lambda(y-\theta)}]^i}{\lambda i}, \qquad (12)$$

and its variance

$$\mathrm{Var}_n\,(\Delta y) = \sum_{i=1}^{n} \frac{2E_i(\Delta y)}{\lambda_i} - E_n^2(\Delta y). \tag{13}$$

Time Evolution and Optimization

The search process proceeds over time, improving technology whenever possible The present value of the system of production and technological research is

$$V(y) = E\left\{\sum_{t=0}^{\infty} [y_t - c(n_t)]\alpha^t\right\}, \tag{14}$$

where y is the present technology level (at $t = 0$) Let $V^*(y)$ denote the value of a system when an *optimal* research policy is followed, it can be written in the form of a recursion functional as

$$V^*(y) = \max_n \left[y - c(n) + \alpha \int_y^\infty V^*(z)\, dF^n(z) + \alpha F^n(y) V^*(y) \right] \tag{15}$$

The last two terms on the right-hand side of (15) are the benefits $B(y, n)$ due to the present experiment

$$B(y, n) = \alpha \int_y^\infty V^*(z)\, dF^n(z) + \alpha F^n(y) V^*(y),$$

and they are, respectively, the expected value of a system starting next period from a better than current technology—if such a technology is found—and the value of the current system weighted by the probability that the outcome of the experiment will not exceed the current yield

Incremental benefits from increasing the number of trials in an experiment are positive and diminishing

$$\begin{aligned} D_nB &\equiv B(y, n) - B(y, n - 1) \\ &= \alpha \int_y^\infty V^*(z)[nfF^{n-1} - (n - 1)fF^{n-2}]\, dz \\ &\quad + \alpha V^*(y)(F^n - F^{n-1}) \end{aligned} \tag{16}$$

Integrating by parts, we get

$$D_nB = \alpha \int_y^\infty \frac{\partial V^*}{\partial z} (F^{n-1} - F^n)\, dz \geq 0, \tag{17}$$

since $(\partial V^*/\partial z) \geq 0$, as a discovery of a better technology can only improve the value of the system This proves that incremental benefits are positive, they are diminishing,

$$D_nB - D_{n-1}B = \alpha \int_y^\infty \frac{\partial V^*}{\partial z} (2F^{n-1} - F^n - F^{n-2})\, dz \leq 0, \quad (18)$$

since $2F^{n-1} \leq F^n + F^{n-2}$

The cost function $c(n)$ is, by assumption, increasing with n at an increasing rate

$$D_nc \equiv c(n) - c(n-1) > 0, \quad (19)$$

$$D_nc - D_{n-1}c > 0$$

Since incremental returns are decreasing, optimal n will be such that

$$|D_{n-1}c - D_{n-1}B| \geq |D_nc - D_nB| \leq |D_{n+1}c - D_{n+1}B| \quad (20)$$

This is illustrated in figure 2 [3]

Economic Properties of Optimal Solutions

1 The optimal number of experiments is a decreasing function of the rate of interest, as D_nB in (16) decreases with r

Since the extent of experimentation, technological research, is a decreasing function of the rate of interest, technological progress will also be a decreasing function of the rate of interest This is one aspect in which the present model resembles Solow's embodiment model (Solow 1960) In our model, technical progress is the outcome of investment in research, in Solow's model it is embodied in new capital assets In both cases, the rate of progress is a function of the rate of interest

2 It is useful to view technological research as filling a gap between basic knowledge and the level of technology in practice (Nelson and Phelps 1966) In the exponential distribution (3), the level of basic research can be represented by the parameter θ The *smaller* the difference $y - \theta$, the *larger* the technological gap and, in our model, the easier it is to improve the technology The analytical difficulty here is that a change in θ modifies the distribution searched and affects $V^*(y)$ by changing all

---

[3] The present optimization framework can undoubtedly be expanded to include adaptive search, sequential experimentation with various stopping rules, and similar cases This has not been done here in order to preserve simplicity and shortness The qualitative results and the major characteristics of research systems focused upon in the present model would have remained unchanged Note in this respect Rothschild's (1974, p 689) conclusion that "not invariably, but in many instances, the qualitative properties of the optimal-search strategies . are the same [in the adaptive case] as in the simpler case when the distribution is assumed known."

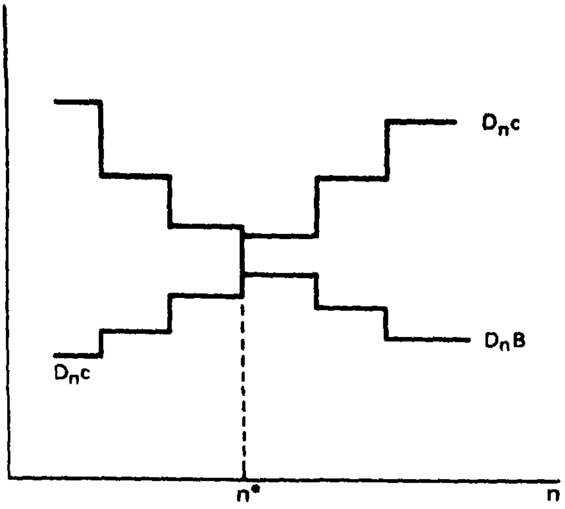


FIG 2 —Optimal n (n^*)

future (expected) steps We therefore show explicitly the effect of a change in θ on the expected value of a system in which *only one* experiment is conducted The benefits of such a system are

$$B^1(y, n) = \frac{\alpha}{r}\int_y^{\infty} znf(z)F^n(z)\, dz + \frac{\alpha}{r}F^n(y)\,y, \tag{21}$$

and in the exponential distribution [with the density function $\lambda e^{-\lambda(z-\theta)}$]

$$\frac{\partial B}{\partial \theta} = -\frac{\partial B}{\partial z},$$

therefore

$$\frac{\partial B^1}{\partial \theta} = -\frac{\alpha}{r}\int_y^{\infty} z\frac{\partial^2 F^n}{\partial z^2}\, dz - \frac{\alpha}{r}\frac{\partial F^n}{\partial z}y$$

Integrating by parts

$$\frac{\partial B^1}{\partial \theta} = \frac{\alpha}{r}[1 - F^n(y)] \geq 0 \tag{22}$$

Since (22) is true for any value of y, current and future, an increase in θ will increase the value of $B(y, n)$ for any unrestricted system

3 Economies of scale can be introduced by multiplying V in (14) and (15) by a scale factor The larger this factor, the higher optimal n and the faster technological progress

This formulation will represent cases in which technology is identical throughout the volume of operation of a firm or an economy, for example, if a new variety will raise yield equally in all fields or if a new chemical compound will increase productivity equally in all its applications. In other cases, however, scale represents, at least partially, variability in the conditions of production the larger the area sown to a crop, the more variable growing conditions are, and this variability factor dampens the direct, proportional scale effect

4 The benefits are a decreasing function of y, since y is the lower bound of the integral in $B(y, n)$ So also optimal number of trials, n, is a decreasing function of the level of technology, y Eventually optimal $n = 0$ and technological research will stop This will be the case when

$$D_n B \leq c(n) \tag{23}$$

for any n Particularly,

$$D_1 B \leq c(1), \tag{24}$$

or, written explicitly,

$$\alpha \int_y^\infty V^*(z) f(z)\, dz + \alpha F(y) V^*(y) \leq c(1) \tag{25}$$

Thus after a certain level, technology will stagnate forever This stagnation will not occur if basic science progresses continuously This possibility is analyzed in the second part of the paper

5 The inequality in (25) is not only the condition for stopping a research project, it is also the condition for not starting one A project will not be undertaken if the expected benefits of even a single trial in a single experiment will not exceed the corresponding cost

$$\begin{aligned} \frac{\alpha}{r} E_1(\Delta y) &= \frac{\alpha}{r} \frac{1}{\lambda} [1 - 1 - e^{-\lambda(y-\theta)}] \\ &= \frac{\alpha}{r} \frac{1}{\lambda} e^{-\lambda(y-\theta)} \leq c(1) \end{aligned} \tag{26}$$

If the system starts from a position represented by (26), then a change in one of the following three parameters can make technological research justifiable (*a*) A reduction in the rate of interest; (*b*) An increase in the basic knowledge parameter θ, since

$$\frac{\partial E_1(\Delta y)}{\partial \theta} = e^{-\lambda(y-\theta)} \geq 0,$$

(*c*) A decrease in λ, since

$$\frac{\partial E_1(\Delta y)}{\partial \lambda} = -\frac{1}{\lambda} e^{-\lambda(y-\theta)} \left(\frac{1}{\lambda} + y - \theta\right) < 0$$

The last is an increase in variance of the population searched. This is an example of variance increasing research; for instance, the International Rice Research Institute in the Philippines accumulated a collection of 15,000 rice varieties increasing substantially the variance of the rice varieties population searched for genetic material.

Basic Research

In terms of the present model, basic research can be classified into three categories: basic research can shift the mean of the distribution searched, it can change its variance, or it can create new technologies—discover new distributions to search. The following discussion will analyze the first case, the last is treated in the concluding section of the paper.[4]

With *mean shifting basic research*, θ is no longer constant. Its per period growth, due to basic research, is $\Delta\theta$, and *it is assumed throughout the discussion that $\Delta\theta$ is constant*. It is shown below that optimal n converges, if $\Delta\theta$ is constant, to a constant level. But a constant level of applied research will not be limited to optimal system. Quite often a research organization is operated at a certain constant level determined by budgets and political circumstances. In such a case an interesting property emerges.

Property (*steady state*).—If both basic research and technological research proceed at constant rates, a stochastic steady state emerges at which technology in practice improves at a rate equal to the rate of advancement of basic research. Formally, if $\Delta\theta$ = constant, n = constant, then eventually $E_n(\Delta y) = \Delta\theta$.

Proof.—For the analysis of the search process, the state variable is not y but, rather, $\xi = y - \theta$, and we have to show that, if n = constant, $\Delta\theta$ = constant, then ξ = constant.

Rewrite $E_n(\Delta y)$ as

$$E_n(\Delta y) = \sum_{i=1}^{n} \frac{1 - (1 - e^{-\lambda\xi})^i}{\lambda i},$$

now

$$\frac{\partial E_n(\Delta y)}{\partial\xi} = - \sum_{i=1}^{n} e^{-\lambda\xi}(1 - e^{-\lambda\xi})^{i-1} \tag{27}$$

As ξ grows, $E_n(\Delta y)$ decreases, and vice versa. There exists a steady-state value of ξ, ξ^*, a function of n and $\Delta\theta$, which will be maintained (stochastically), it is defined by

$$E_n(\Delta\xi^*) = \sum_{i=1}^{n} \frac{1 - (1 - e^{-\lambda\xi^*})^i}{\lambda i} - \Delta\theta = 0 \tag{28}$$

Q.E.D.

---

[4] The case of variance increasing research is discussed, in a similar framework, in Evenson and Kislev (1975, chap. 8).

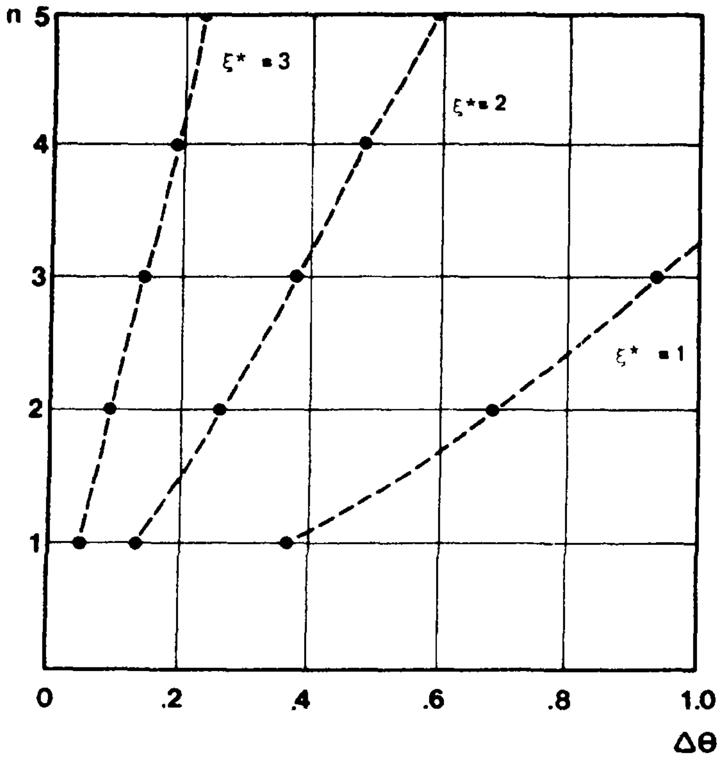


FIG 3 —Steady-state ξ for $\lambda = 1$, $n = 1, 2, \quad, 5$, $0 < \Delta\theta < 1$

Steady-state ξ can be written as

$$\xi^* = \xi^*(n, \Delta\theta) \tag{29}$$

Figure 3 illustrates equation (29) for $\xi^* = 1, 2, 3$, $\lambda = 1$, $n = 1, 2, \quad, 5$, $0 < \Delta\theta < 1$ The broken lines in the figure only connect steady-state $(n, \Delta\theta)$ pairs, other points on the lines are meaningless We shall now investigate algebraically the properties of (29)

For a given n, steady-state ξ is a decreasing function of $\Delta\theta$ To see this, differentiate (28) with respect to $\Delta\theta$

$$\frac{\partial E_n(\Delta\xi^*)}{\partial\Delta\theta} = -\frac{\partial\xi^*}{\partial\Delta\theta}\sum_{i=1}^{n}(1 - e^{-\lambda\xi^*})^{i-1}e^{-\lambda\xi^*} - 1 = 0; \tag{30}$$

$$\frac{\partial\xi^*}{\partial\Delta\theta} = -\left[\sum_{i=1}^{n}(1 - e^{-\lambda\xi^*})^{i-1}e^{-\lambda\xi^*}\right]^{-1} \tag{31}$$

The interpretation of this result is that in the steady state for a given level of technological research (a given n), the higher the rate of basic knowledge, the larger the gap between basic knowledge and technology in practice (the smaller ξ^*). This property stems from the fact that the bigger the gap, the more productive will technological research be. In the steady state this gap, represented by the variable ξ^*, is the only endogenous variable in the system; the bigger $\Delta\theta$, the larger the gap which will maintain $E_n(\Delta y) = \Delta\theta$.

Steady-state ξ can even become negative. For example, if $\lambda = 1$ and $n = 1$, $\xi^* < 0$ for $\Delta\theta < 1.0$. This can readily be seen by rewriting (28) for $\Delta\theta = 1$:

$$E_1(\Delta\xi^*) = e^{-\xi^*} - 1 = 0 \qquad (28')$$

Clearly, $\xi^* = 0$ in (28′) and by (31) $\xi^* < 0$ for $\Delta\theta > 1$.

A negative ξ^* means that basic knowledge has advanced, relative to technology in practice, to such a level that *any search* will find (with probability one) superior technologies. It should not be surprising that such cases are not observed often in the real world. Technological research is an economic activity, perhaps not always conducted optimally, but surely directed into areas with obvious and safe gains and, thereby, eliminating these potential gains.

To find the behavior of ξ^* as a function of n, approximate the sum in (28) by an integral:

$$E_n(\Delta\xi^*) = \int_0^n \frac{1 - (1 - e^{-\lambda\xi^*})^i}{\lambda i}\, di - \Delta\theta = 0 \qquad (28'')$$

Differentiating (28″) with respect to n and rearranging terms, we get

$$\frac{\partial\xi^*}{\partial n} = \frac{1 - (1 - e^{-\lambda\xi^*})^n}{n\lambda e^{-\lambda\xi^*} \int_0^n (1 - e^{-\lambda\xi^*})^{i-1}\, di} > 0 \qquad (32)$$

The steady-state ξ grows with n—the more technological research is conducted, the smaller the steady-state gap between technology and basic knowledge.

To find the optimal level of experimentation, note that the present value of the system in a steady state is

$$\begin{aligned} \omega &= \sum_{t=0}^{\infty} [E(y_t) - c(n)]\alpha^t \\ &= \frac{y_0}{r} + \sum_{t=0}^{\infty} \left[\sum_{t=0}^{t} E_n(\Delta y) \right] \alpha^t - \frac{c(n)}{r} \\ &= \frac{\theta_0 + \xi^* - c(n)}{r} + \Delta\theta \sum_{t=0}^{\infty} t\alpha^t. \end{aligned} \qquad (33)$$

The only terms in the last line of (33) which are functions of n are ξ^* and $c(n)$ Therefore, the problem of optimal n can be formulated, disregarding the other terms, as

$$\operatorname*{Max}_{n}\ [\xi^*(n, \Delta\theta) - c(n)] \tag{34}$$

The maximization in (34) is not limited to $t = 0$, as one may erroneously conclude from an examination of (33), any t can be viewed as $t = 0$ The problem is to find what constant level of technological research to maintain so as to reach an optimal (stochastically) constant level of ξ^* In other words, in the steady-state situation, technological research does not determine the *rate of progress* of technology—this rate is equal to the exogenously determined rate of progress of basic knowledge—rather, technological research determines the *level* of technology in practice or, rather, the constant technological gap [5] In this respect, the present model is similar in character to the Nelson and Phelps (1966) model, where in a steady-state equilibrium schooling determines the technological gap, while the rate of technical progress is determined by the rate of advancement of basic science

From (32) it can be shown that

$$\frac{\partial^2\xi^*}{\partial n^2} < 0, \tag{35}$$

and since $D_n c \equiv c(n) - c(n - 1) \geq 0$, optimal n will be the value for which

$$|D_{n-1}\xi^* - D_{n-1}c| \geq |D_n\xi^* - D_n c| \leq |D_{n+1}\xi^* - D_{n+1}c|, \tag{36}$$

where $D_n\xi^* \equiv \xi^*(n, \Delta\theta) - \xi^*(n - 1, \Delta\theta)$

What effect will the rate of basic research have on optimal n in the steady state? On the one hand, the faster the advancement of basic knowledge, the higher the productivity of applied research, on the other hand, in the steady state, the higher the level of basic research, the smaller the technological gap In fact, as can be seen by differentiating (32) with respect to $\Delta\theta$, the sign of $\partial^2\xi^*/\partial n\, \partial\Delta\theta$ is indeterminate, and the magnitude of this second derivate is very small In the numerical example used in the illustration of figure 3, optimal n, over the range $0 < \Delta\theta < 1$, is almost always independent of $\Delta\theta$ This can be realized by examining figure 4, where $D_n\xi^*$ is plotted against n for $\Delta\theta = 1, 5, 9$ Note, for example, that if $c(n) = 4n$, $(D_n c = 4)$, optimal n will be $n = 3$ for any $\Delta\theta$ at least over the range $1 \leq \Delta\theta \leq 9$

This tendency of the applied research to be a decreasing (at least, not

---

[5] An exception is the case of $n = 0$—no technological research—in which technology in practice is stagnant and the technological gap constantly increases as basic research adds to the stock of basic knowledge

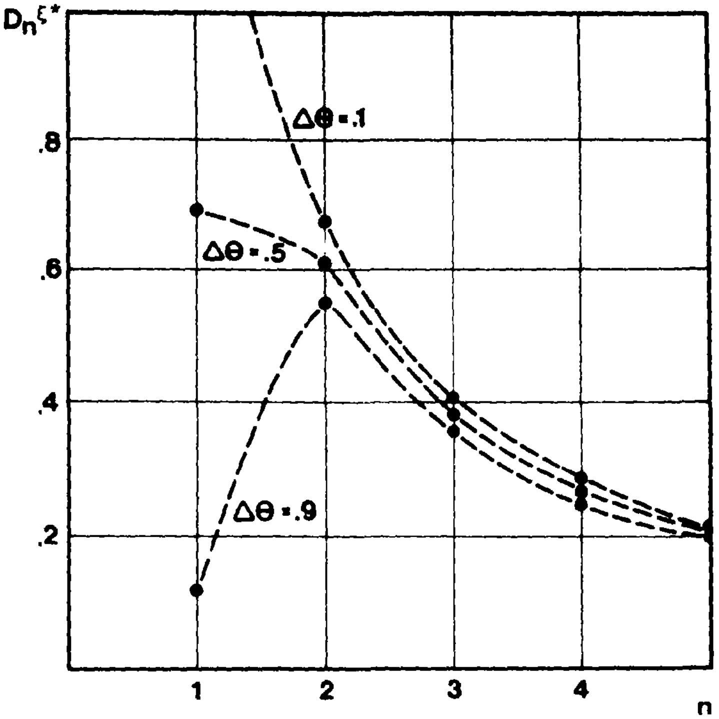


FIG 4 —$D_n\xi^* \equiv \xi^*(n, \Delta\theta) - \xi^*(n - 1, \Delta\theta)$ for $\lambda = 1, n = 1, 2, \ldots, 5$

an increasing) function of the rate of basic scientific advancement is a reflection of the steady-state conditions, where the faster basic knowledge accumulates, the larger the technological gap (the smaller ξ^*) and the smaller may be the optimal n A different case is discussed in the following section

Exogenously given $\Delta\theta$ and the cost function $c(n)$ determine the pair optimal n and ξ^* in the steady state If the system starts with $\xi < \xi^*$, then marginal returns to experimentation will be larger than in the steady state and optimal n will tend to be larger than steady-state optimal n With time, as $\xi \to \xi^*$, optimal n will approach its steady-state value. A similar convergence to equilibrium, in the opposite direction, will take place if initially $\xi > \xi^*$ The steady state is the long-run equilibrium stable position of an optimal system

Concluding Remarks on the Discovery of New Technologies

Science creates a great variety of applicable technologies, from electricity to antibiotics, from atomic energy to genetic-engineering. In terms of the present model, the discovery of new technologies is represented by the opening up of new exponential distributions for search. For simplicity, assume that all technologies are characterized by identical distributional attributes: identical λ and θ parameters; moreover, assume that $\theta = 0$ in all distributions These are strong assumptions, but they permit focusing the discussion on some essential issues The relaxation of these assumptions will complicate the analysis greatly and will have to be postponed to another occasion

Let k be the technology index, k is a vintage parameter The more recent the technology, the higher its k value The expected income from an experiment in technology k is

$$E_{n(k)}[\Delta y(k)] = \sum_{i=1}^{n(k)} \frac{1 - [1 - e^{-\lambda \xi(k)}]^i}{\lambda i} \tag{37}$$

For newly discovered technologies, (37) reduces to

$$E_{n(k)}[\Delta y(k)] = \sum_{i=1}^{n(k)} \frac{1}{\lambda i}, \tag{37'}$$

since $y = \theta$ for new technologies

At any point in time there exists a spectrum of technologies in the economy Technological research is now an industry with a rising supply function

$$c(n),\ n = \sum_{k} n(k),$$

of factors of research Experiments will be allocated to technologies according to their comparative standing in the current technological spectrum

The discussion that led to equation (20) can now be repeated, and again that equation is the condition for optimal allocation of experiments with one departure from the previous analysis If there is a large number of distribution to search, then $D_n c$ is now the incremental cost function for the industry, it is taken as datum when considering a single distribution In other words, with a very large number of technologies, the allocation of research efforts becomes analogous to the organization of a competitive industry with separate firms doing research in the separate distributions.

In addition to ordering distributions by their vintage index k, they can now be ordered by their $y(k)$ value There will be a strong (negative) correlation between y and the vintage index, but these will not be the same orderings A break-even value of y will be y^* maintaining

$$\frac{\alpha}{r}\frac{1}{\lambda}e^{-y^*} = c(1) \tag{38}$$

The term on the left in (38) is identical to the present value of the incremental income due to an experiment of one trial. Distributions with $y > y^*$ will not be subject to search any more These distributions are *technologically exhausted* Exhaustion is an economic phenomenon, a reduction in the cost function, for example, will cause the resumption of search in previously exhausted technologies

Technological exhaustion should be carefully distinguished from *obsolescence* The last is a market phenomenon, reflecting changing comparative advantages of technologies and not treated in the present study.

The rate of progress of basic science is measured in the model by the number of new technologies discovered per time period. The larger this number, the higher the demand for research factors and the higher the rate of experimentation At the same time, the higher the rate of advancement of basic science, the faster the rate of exhaustion of old technologies In fact, in a steady state, the number of exhausted technologies will be equal to the number of newly discovered technologies, and the larger this number, the *less* search will be conducted in a typical technology before its exhaustion

References

Binswanger, H P "A Microeconomic Approach to Induced Innovation" *Econ J* 84 (December 1974) 940–58

De Groot, Morris H *Optimal Statistical Decisions* New York McGraw-Hill, 1970.

Epstein, B "Elements of the Theory of Extreme Values" *Technometrics* 2 (February 1960) 27–41

Evenson, Robert E, and Kislev, Yoav *Agricultural Research and Productivity* New Haven, Conn Yale Univ Press, 1975

Gumbel, E J *Statistics of Extremes* New York Columbia Univ Press, 1958.

Karlin, Samuel "Stochastic Models and Optimal Policy for Selling an Asset" In *Studies in Applied Probability and Management Science*, edited by K J Arrow, S Karlin, and H Scarf, Stanford, Calif Stanford Univ Press, 1962

Kohn, Meir G, and Shavell, Steven. "The Theory of Search" *J Econ Theory* 9 (October 1974) 93–123

Nelson, R R "Uncertainty, Learning and the Economics of Parallel Research and Development Efforts" *Rev Econ and Statis* 43 (November 1961)· 351–64

Nelson, R R, and Phelps, Edmund S "Investment in Humans, Technological Diffusion, and Economic Growth." *A E R* (Papers and Proceedings) 56 (May 1966) 69–75

Rothschild, Michael "Searching for the Lowest Price When the Distribution of Prices Is Unknown" *J P E* 82 (July/August 1974) 689–711

Solow, Robert M "Investment and Technical Progress" In *Mathematical Methods in the Social Sciences*, edited by Kenneth J Arrow, Samuel Karlin, and Patrick Suppes Stanford, Calif Stanford Univ Press, 1960

Stigler, G J. "The Economics of Information." *J.P E* 69 (June 1961) 213–15.

The Marginal Cost of Public Funds

Edgar K. Browning

University of Virginia

The marginal cost of public funds is the direct tax burden plus the marginal welfare cost produced in acquiring the tax revenue. This paper estimates that the marginal cost of public funds for taxes on labor income in the United States ranges from $1.09 to $1.16 per dollar of tax revenue, depending on the progressivity of the change in the tax structure. Thus, government expenditures must be at least 9–16 percent more productive than private expenditures to produce a net welfare gain. In addition, the total welfare cost of income taxes in 1974 is estimated at $19 billion.

Taxes introduce distortions in the allocation of resources, and much theoretical and empirical work in economics has been devoted to analyzing and estimating the welfare costs (or excess burdens) of the variety of taxes in existence. Very little attention has been given, however, to the significance of the welfare cost of taxation for the analysis of public expenditure programs. If the financing of expenditure programs involves a welfare cost, then this cost should be considered part of the opportunity cost of the expenditure programs. Put briefly, when the government spends $100, the opportunity cost is $100 plus the additional welfare cost involved in acquiring the funds. Thus, an expenditure program will be efficient only if its benefits exceed the direct tax cost by an amount at least as large as the additional welfare cost of the funds.

A number of economists have, of course, recognized the importance of considering (in some way) the welfare cost of taxation in analyzing government expenditures. Pigou (1947) mentioned this point many years ago, and more recently Buchanan and Tullock (1965), Vickrey (1963),

---

Research on this paper was conducted while the author was a sesquicentennial associate of the Center for Advanced Studies at the University of Virginia.

[*Journal of Political Economy*, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved.

and Johnson and Pauly (1969) have discussed this aspect of tax-induced distortions In addition, in his survey of cost-benefit analysis, Musgrave (1969) has suggested that Harberger's estimates of the average welfare costs of taxes on labor and capital should be added to the direct costs of taxes [1] The present paper builds on the insights of these works by examining the problem of measuring the marginal cost of public funds Section 1 develops the basic theory for an economy using a tax on labor income to finance government expenditures In Section 2 estimates for taxes on labor income in the United States tax system are developed The results suggest that the marginal cost of public funds is between \$1 09 and \$1 16 per dollar of tax revenue

1. The Marginal Welfare Cost of Income Taxation

We shall assume an economy with a competitive private sector where the government uses a tax on labor income to finance its expenditure programs Taxes on labor income distort the labor supply decisions of workers, and the problem is to determine exactly how this distortion should be incorporated into an analysis of government expenditures It will be assumed that the expenditure programs themselves do not distort labor supply decisions at the margin—that is, they can be viewed as lump sum transfers [2] This simplifying assumption allows us to concentrate on the tax-induced distortions and ignore possible differential effects resulting from the specific way the tax revenue is used

A tax on labor income with a marginal tax rate of m_t lowers a worker's net wage rate from w, the market wage rate, to $w(1 - m_t)$ As a consequence, the quantity of labor services supplied will be below the quantity the worker would supply with a tax which did not distort his labor supply decision Figure 1 illustrates this situation The worker's supply curve is S, and is drawn to include only substitution effects of wage rate changes With a marginal tax rate of m_t, L_2 units of labor are supplied By supplying L_2 units rather than L_1 units the worker sacrifices earnings of BAL_1L_2 and gains leisure time valued at DAL_1L_2 The difference, area BAD, is the total welfare cost of the tax

If the government were to implement a new expenditure program, or expand an existing one, it must raise additional revenue Suppose the

---

[1] Musgrave errs, however, in suggesting the use of average welfare cost per dollar of revenue since it is the marginal welfare cost that is relevant for cost-benefit analysis As will be shown below, marginal welfare costs of taxes are typically much larger than average welfare costs

[2] In some cases, particularly those involving redistribution, this assumption would not be valid. For example, when the income tax is used to finance transfers of a negative income tax variety, the marginal tax rates of both taxpayers and recipients will rise. This distorts the labor supply decisions of both groups, and an analysis of such a policy should ideally evaluate both distortions I have attempted to do this in Browning (1975*b*).

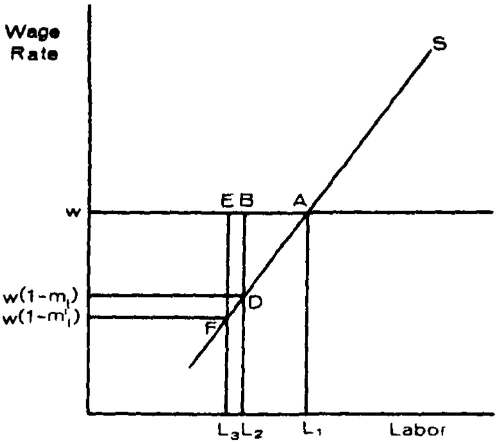


FIG. 1

revenue requirements necessitate an increase in the marginal rate from m_i to m'_i. Then the marginal welfare cost of this increment in the tax rate is the area *BEFD*. Thus, the cost associated with the new expenditure program is clearly the direct burden of the higher tax rate—the amount of revenue raised—plus the added distortion, area *BEFD*. (In addition, there may be additional administrative and compliance costs, but they will be ignored in this section.) If economic efficiency is the criterion for judging the expenditure program, then its benefits must exceed the required revenues plus area *BEFD*. Clearly, neither the total welfare cost nor the average welfare cost of the tax is really relevant in evaluating expenditure programs; instead it is the marginal welfare cost that is critical.

The size of the marginal welfare cost is, then, of great importance in the evaluation of government expenditures. Harberger's pioneering work in this area provides the basis for evaluating this cost for income taxation. Harberger (1964) shows that the total welfare cost for an individual worker can be expressed as

$$W_i = \tfrac{1}{2}\eta(m_i)^2 Y_i, \tag{1}$$

where W_i is the total welfare cost, m_i is the marginal tax rate, η is the elasticity of labor supply (expressing the substitution effect alone), and Y_i is gross labor income, or earnings.[3] Now let us assume that the tax is a proportional tax on labor income (m_i is thus the same for all taxpayers and is equal to the average tax rate), and that the elasticity of labor supply is the same for all workers Then the total welfare cost aggregated over all workers is

---

[3] This expression measures only the welfare cost due to the distortion in the worker's decision concerning hours of work. An income tax also distorts occupational choice by inducing a substitution in favor of jobs with nonpecuniary advantages. Thus, estimates of welfare costs using equation (1) will be biased downward to some degree.

$$W = \tfrac{1}{2}\eta m^2 \sum Y_i, \tag{2}$$

where $\sum Y_i$ is total labor income[4]

With a proportional tax, total tax revenue, T_p, can be expressed as

$$T_p = m \sum Y_i \tag{3}$$

Assuming that the tax base, $\sum Y_i$, is not affected by a small change in the tax rate, the additional welfare cost produced by a change in the tax rate is

$$dW = \eta m \sum Y_i \, dm, \tag{4}$$

and the additional revenue produced is

$$dT_p = \sum Y_i \, dm \tag{5}$$

The marginal welfare cost per dollar of revenue is then equation (4) divided by equation (5), or

$$dW/dT_p = \eta m \tag{6}$$

The marginal cost of public funds is the marginal welfare cost of taxation plus the direct cost, or simply $\eta m + 1$[5] Note that the marginal welfare cost is greater than the average welfare cost The average welfare cost is equal to W/T_p, or $\tfrac{1}{2}\eta m$, exactly half the marginal welfare cost

To get some feeling for the likely quantitative importance of the marginal welfare cost of income taxation, let us assume some plausible values for η and m On the basis of recent empirical work, it appears that

---

[4] The summation of individual welfare costs as given by (1) to derive the total welfare cost as given by (2) is not fully appropriate unless the marginal product of labor does not decline with a simultaneous increase in hours of work by all workers In other words, expression (2) is valid only if the aggregate demand curve for longer hours of work per worker is perfectly elastic. If this is not true, the correct formula is $\tfrac{1}{2}[\eta \varepsilon_d/(\eta + \varepsilon_d)]m^2 \sum Y_i$, where ε_d is the elasticity of this demand curve Estimates based on equation (2) when demand is less than perfectly elastic are therefore biased upward However, for low values of η (as will be assumed in this study), this bias is not very large For example, with $\eta = 0\ 20$ and $\varepsilon_d = 2\ 0$, equation (2) will overestimate the true welfare cost due to labor supply distortions by only 10 percent In addition, there is an opposite bias involved in using equation (2), as noted in the previous footnote, so the net bias is not clear

[5] In this derivation it was assumed that the tax base did not change in response to a change in the tax rate. If this assumption is dropped, the marginal welfare cost equals $\eta m/\varepsilon + 1$, where ε is the elasticity of the tax base with respect to a change in the tax rate. Thus, equation (6) will underestimate the marginal welfare cost if the tax base falls ($\varepsilon < 0$) and overestimate the marginal welfare cost if it rises ($\varepsilon > 0$) when the tax rate is increased. Since there is no empirical basis for assigning a value to ε, it will simply be assumed that it is zero in this paper

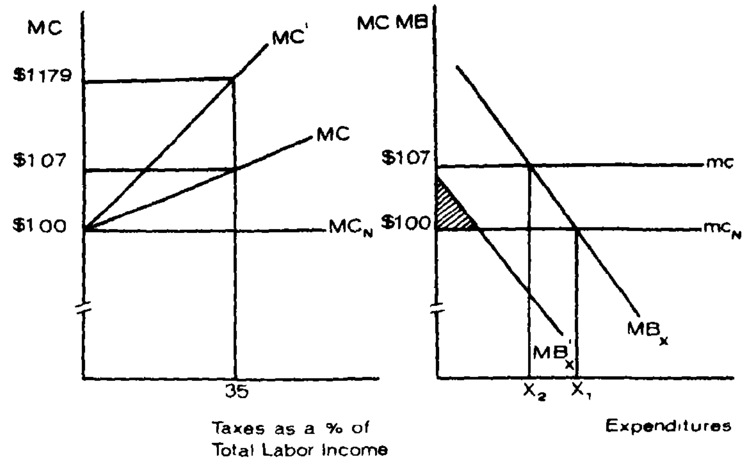


FIG. 2

a value for η of 0.20 is a fairly conservative estimate.[6] Since total taxes as a percentage of net national product are about 35 percent in the United States, a value for m of 0.35 will be used. Given these values, the marginal welfare cost is 7 percent of tax revenue, and hence the marginal cost of public funds is 107 percent, or $1.07 per dollar of revenue. Thus, government expenditures would have to be 7 percent more beneficial (at the margin) than private expenditures for the programs to constitute a net welfare gain.

Figure 2 can conveniently summarize the analysis to this point. In the left-hand panel, the marginal cost of public funds is measured vertically and taxes as a percentage of total labor income measured horizontally. The MC curve gives the marginal cost of public funds for a flat rate tax on total labor income. At a tax rate of 35 percent, the marginal cost per dollar of additional revenue is equal to $1.07, as just indicated. The vertical difference between MC and MC_N (the marginal cost of public funds with nondistorting taxes) is the marginal welfare cost per dollar of tax revenue.

The right-hand panel of figure 2 shows how the marginal cost of public funds should be incorporated into an analysis of expenditure programs. The central point here is that the marginal cost of public funds for any specific expenditure program is constant at a level determined by the overall weight of taxes in the economy. Thus, when $m = 0.35$, the marginal cost for *all* expenditure programs is \$1.07 since we may assume that no individual program is large enough to affect the marginal tax rate significantly, the marginal tax rate reflects the combined effect of all tax-financed expenditures, and it is imperceptibly affected by any single expenditure. (For an appropriate analogy, the MC curve may be likened to an upward sloping industry supply curve, while the mc curve is akin to the horizontal supply curve confronting an individual purchaser.)

---

[6] See Bloch (1973) and Cain and Watts (1973). Most recent research has emphasized that labor supply elasticities vary among demographic groups. Insofar as this is true, η should be interpreted as a weighted average of the elasticities of different groups. Interpreted in this way, I think the assumption of a value of 0.20 is quite conservative.

If the marginal benefits of spending on program X are given by MB_X, then the efficient level of expenditures with a proportional income tax is X_2, a lower level than if expenditures were financed with distortionless taxes (X_1). If the marginal benefit curve is MB'_X, then the efficient level of expenditures is zero. Thus, even though there is a potential gain from government spending (given by the shaded area), that gain cannot be realized if the distorting income tax must be used as a method of finance. This conclusion has immediate relevance for the various "market failure" (e.g., externality) arguments for government expenditures. What the present analysis shows clearly is that the presence of a market failure does not necessarily imply that any expenditure program exists which will produce a net welfare gain. Instead, the magnitude of the market failure must exceed some critical value (as determined by the marginal cost of public funds) before an expenditure program can be capable of achieving a net gain.

We may now consider how the use of nonproportional income taxes affects the analysis. As an illustration, we will use the simplest form of progressive taxation—a degressive tax. A degressive tax is a flat rate tax above an exemption. Thus, the marginal tax rate is the same for all taxpayers, but average tax rates rise with total income, and the tax is progressive overall. With such a tax, the total welfare cost is given by

$$W = \tfrac{1}{2}\eta m^2 \sum_{n} Y_i, \qquad Y_i > K, \tag{7}$$

where $\sum_n Y_i$ is the total income of those subject to the tax, and K is the amount of exempted income per taxpayer. Total tax revenue is:

$$T_D = m \sum_{n} (Y_i - K), \qquad Y_i > K \tag{8}$$

Differentiating (7) and (8) and dividing yields the marginal welfare cost per dollar of revenue

$$dW/dT_D = \eta m \frac{\sum_n Y_i}{\sum_n (Y_i - K)} \tag{9}$$

A comparison of equations (9) and (6) shows that the marginal welfare cost of a degressive tax is greater than that of a proportional tax yielding the same total revenue. There are two reasons for this. First, the marginal tax rate must be higher for the degressive tax when both taxes yield the same total revenue since the tax base of the degressive tax is smaller. (The marginal tax rate of the degressive tax will be equal to

$$\frac{\sum_n Y_i}{\sum_n (Y_i - K)}$$

times the marginal tax rate of the proportional tax.) Second, the increment in the marginal tax rate required to raise additional revenue is greater for the degressive tax because of the smaller tax base. This is reflected in (9) by the term

$$\frac{\sum_n Y_i}{\sum_n (Y_i - K)},$$

which is always greater than one. These two factors together imply that the marginal welfare cost of the degressive tax can be expressed as the marginal welfare cost of the proportional tax times

$$\left[\frac{\sum_n Y_i}{\sum_n (Y_i - K)}\right]^2$$

The squared term always exceeds one for a degressive tax, and will be greater the larger is the exemption.

Even with a seemingly small exemption, the marginal welfare cost of a degressive tax will be substantially above that of a proportional tax. If, for example, the exemption is equal to 40 percent of average income, the tax base of the degressive tax would be about 62.5 percent of the tax base of the proportional tax, and

$$\frac{\sum_n Y_i}{\sum_n (Y_i - K)}$$

would equal 1.6.[7] To raise 35 percent of total income in taxes would then require a marginal tax rate of 56 percent. Inserting these values into (9) and solving yields a marginal welfare cost of 17.9 percent of tax revenue. (The marginal cost of public funds with the degressive tax is shown by *MC'* in fig. 2.) Thus, the marginal welfare cost of this degressive tax is 2.5 times larger than that of a proportional tax yielding the same revenue. The important implication of this for expenditure policy is that expenditures should be smaller with a degressive income tax than with a proportional income tax. In general, the more progressive the tax the greater is the marginal cost of public funds, and the lower is the efficient volume of government expenditures.

To this point we have considered taxes that impose the same marginal tax rate on all taxpayers When a tax with a graduated rate structure is employed each rate bracket is subjected to a different marginal tax rate, and each bracket must then be considered separately. A more difficult problem posed by graduated rate taxes arises because additional revenue can be raised by many different changes in the rate structure, and each possible change may correspond to a different marginal welfare cost Unlike the proportional and degressive taxes, there is no rule inherent in a graduated rate tax that specifies how changes in revenue are to be accomplished Of course, for any specified method of raising additional revenue the marginal welfare cost can be estimated, as will be seen in the next section

---

[7] This is a rough figure calculated from data in U.S. Bureau of the Census (1973, table 2).

The analysis in this section has provided some insight into the concept of marginal welfare cost and its relation to the income tax structure In the next section we will derive estimates of the marginal welfare cost (and, hence, the marginal cost of public funds) for the United States income tax structure

2. Application to the U.S. Tax System

A major difference between the U S tax system and the model used in the last section is that there are several different taxes that are levied on labor income in the United States The federal individual income tax is obviously the most important, but the social security payroll tax, state income taxes, and other taxes also distort labor supply decisions in the same way This fact necessitates a slight modification in the analysis developed in the last section If we have two taxes which are levied on labor income at marginal rates of m_1 and m_2, the total welfare cost for an individual is

$$W_i = \tfrac{1}{2}\eta(m_1 + m_2)^2 Y_i \qquad (10)$$

Thus, the total welfare cost, and the marginal welfare cost, will depend on the effective marginal tax rate on income, $m_1 + m_2$ in the above example In this setting it is ambiguous to speak of the total welfare cost of one tax alone since $\tfrac{1}{2}\eta(m_1)^2 Y_i + \tfrac{1}{2}\eta(m_2)^2 Y_i$ does not equal the total welfare cost of the two taxes together However, the total welfare cost of both taxes together can be estimated by (10), and the marginal welfare cost of a change in either tax can also be estimated without ambiguity

To estimate the marginal cost of public funds it is necessary to have the set of effective marginal tax rates implicit in the tax system and the total labor income subject to each effective rate Total wage and salary income by marginal tax bracket in the federal individual income tax is available from the Treasury tax model [8] We require in addition the increments in marginal tax rates for each federal bracket due to the other taxes on labor income so these increments can be added to the federal rates to arrive at the effective marginal tax rates. There are four major taxes other than the federal income tax that distort labor supply decisions social security payroll taxes,[9] state and local sales taxes, state and local income taxes, and excise taxes.[10] These taxes raised $148 billion in revenue in 1973, and this represented an *average* tax rate of 18 8 percent on total employee compensation (Council of Economic Advisers 1975) It would, however, be an error to simply interpret this 18 8 percent figure as the effective marginal increment for every federal tax bracket

---

[8] I would like to thank Frederick Hickman and Walter Stromquist of the Treasury Department for making this information available.

Three important factors suggest that the increment in marginal tax rates due to these other taxes will vary among the federal tax brackets First, the social security payroll tax applies only to the first $14,100 of earnings of each worker Therefore, for workers earning above this ceiling the marginal rate of the payroll tax is zero [11] Second, most state and local income taxes are highly progressive Third, state and local taxes are deductible in computing federal tax liability. This third factor makes the effective marginal tax rate less than the sum of the federal and state-local rates For example, a person in a 50 percent federal bracket and a 10 percent state income tax bracket is subject only to an effective marginal tax rate of 55 percent because of this deductibility provision in the federal law

These factors suggest that, on balance, the increments in marginal tax rates due to these taxes will be lower the higher the federal tax rate Table 1 shows estimates of the incremental marginal tax rates in column 2 for each federal tax bracket (column 1) [12] (Note that the highest federal rate is 50 percent due to the recently enacted maximum tax on earned

---

[9] It is assumed that social security benefits are not closely related to previous taxes paid If they are, social security taxes may produce no labor supply distortion On this point, see Browning (1975*a*)

[10] Excise taxes are not identical to general taxes on labor incomes, but they do distort the choice between leisure and the taxed goods (and also between taxed and untaxed goods) In theory, an excise tax may lead to greater, less, or the same work effort as an equal yield income tax In the absence of evidence to the contrary, it seems preferable to assume that excises have the same effect on work effort than to ignore them altogether. Ignoring them, however, would make little difference to the estimates made here since they represent only 2 percent of total employee compensation

[11] It would be incorrect to infer that the labor supply decisions of all *families* with earnings above $14,100 are unaffected by the social security tax Not only may such families have multiple earners, one or both of whom are subject to the tax at the margin, but also the labor supply decisions of unemployed family members may be influenced by the tax For example, a wife may choose not to work partly because of the social security taxes she would have to pay on the first $14,100 of her earnings

[12] Interpreting these marginal rates as effective rates of taxation is not fully correct when some uses of income are not subject to tax because of the deductions permitted in the tax law Deductions serve to lower the effective rate of earnings, but they also introduce distortions in the use of income so the net effect on welfare cost is theoretically indeterminate

TABLE 1

TOTAL WELFARE COST BY MARGINAL RATE CLASS IN 1974

Marginal Rate in Federal Income Tax (m_i^F) (%) (1)	Increment Due to Other Taxes (%) (2)	Effective Marginal Tax Rate (m_i) (%) (3)	Wage and Salary Income in Class ($ millions) (4)	Total Welfare Cost by Rate Class ($ millions) (5)	Total Welfare Cost by Rate Class Using (m_i^F) ($ millions) (6)
14	20	34	27,603	319.1	54.1
15	20	35	17,443	213.7	39.2
16	20	36	20,430	264.8	52.3
17	19	36	24,485	317.3	70.8
18	19	37	5,405	74.0	17.5
19	19	38	146,828	2,120.2	530.0
21	18	39	21,782	331.3	96.1
22	18	40	148,265	2,372.2	717.6
23	18	41	3,857	64.8	20.4
24	17	41	21,340	358.7	122.9
25	17	42	121,023	2,134.9	756.4
27	16	43	8,577	158.6	62.5
28	16	44	59,426	1,150.5	465.9
29	15	44	4,688	90.8	39.4
31	15	46	3,858	83.6	37.1
32	14	46	32,469	687.0	332.5
34	14	48	1,705	39.0	19.7
35	13	48	224	5.2	2.7
36	13	49	18,884	453.4	244.8
38	12	50	936	23.4	13.5
39	12	51	10,556	274.6	160.6
40	11	51	1,098	28.6	17.6
41	10	51	67	1.7	1.1
42	9	51	7,066	183.8	124.6
45	8	53	5,446	153.0	110.3
48	7	55	3,403	103.3	78.4
50	6	56	24,842	779.0	621.1
Total			741,706	12,786.8	4,809.1

income.) While these estimates are, admittedly, inexact, I think they are probably reliable enough for our purposes.

Column 3 gives the effective marginal tax rates and column 4 the total wage and salary income in each marginal rate class. Before we turn to the estimation of the marginal welfare cost, it is interesting to use these data to compute the total welfare cost due to labor supply distortions arising from taxes on labor income. Column 5 gives the total welfare cost for each tax bracket; in making these calculations it was assumed, as earlier, that $\eta = 0.20$ and is the same for all rate classes. Summing the figures in column 5 gives the total welfare cost due to labor supply distortions in 1974 $12.8 billion. This figure does not include the welfare cost borne by those who did not pay federal income taxes (but who still paid social security and other taxes), and adjusting for this omission would probably raise the total above $13 billion. In addition, government administrative costs and citizen compliance costs are also elements of the welfare cost that should be added Musgrave and Musgrave (1973, p 460) suggest that (at least at the federal level) these costs are 2–2.5 percent of tax revenue Since the taxes under consideration here produced about $287 billion in revenue in 1974, administrative and compliance costs can be estimated at approximately $6 billion Thus, the total welfare cost of these taxes was in the neighborhood of $19 billion in 1974, or 6 6 percent of tax revenue

The data in table 1 can also be used to demonstrate the importance of using the effective marginal tax rates in the computations Harberger (1964), when examining the welfare cost of the federal income tax, used only the marginal rates of the federal tax in his calculations As mentioned earlier, this procedure is incorrect when there are other taxes on labor incomes, but the results of such a calculation do serve to suggest the importance of considering all taxes together Column 6 uses the marginal rates of the federal income tax to estimate the total welfare cost at $4 8 billion, far below the $12 8 billion estimate when the effective tax rates are used in the calculation Since the total welfare cost varies with the square of the effective marginal tax rate, ignoring even a small part of the effective rate can lead to substantial underestimation of the total welfare cost The "cumulative tax rate" problem that has been emphasized in connection with welfare programs is clearly also of considerable importance in the analysis of tax programs [13]

Turning now to the estimation of marginal welfare costs, we shall consider three changes in the tax structure corresponding to the proportional, degressive, and graduated rate taxes considered in the last section To estimate the marginal welfare cost of a proportional tax that is added to the present system we cannot simply apply equation (6) of the last section since that equation was derived on the simplifying assumption that the initial marginal tax rates were the same for all taxpayers. The appropriate formula for the case when taxpayers are in different initial rate classes, however, is easily derived The total welfare cost is

$$W = \tfrac{1}{2}\eta(m_1)^2 \sum_j Y_{1j} + \tfrac{1}{2}\eta(m_2)^2 \sum_k Y_{2k} + \cdots, \qquad (11)$$

where m_1 is the effective marginal tax rate in the first marginal rate class and $\sum_j Y_{1j}$ is the total labor income in that bracket, and so on for the other brackets Then

$$dW = \eta m_1 \sum_j Y_{1j}\, dm_1 + \eta m_2 \sum_k Y_{2k}\, dm_2 + \cdots, \qquad (12)$$

---

[13] The "cumulative tax rate" effect refers to the fact that recipients of transfers from several different welfare programs, each with its own implicit marginal tax rate, will often be subjected to a very high combined, or effective, tax rate. The same phenomenon occurs for taxpayers who pay several different taxes, and it is the effective marginal rate that is relevant in analyzing the taxpayers' reactions A good discussion of the cumulative tax rate problem in existing welfare programs can be found in Aaron (1973).

TABLE 2

MARGINAL WELFARE COST VARIOUS CHANGES IN TAX STRUCTURE (%)

PROPORTIONAL TAX			DEGRESSIVE TAX		PROGRESSIVE TAX	
m_i (1)	ηm_i (2)	$\eta m_i \frac{\sum_j Y_{ij}}{\sum Y_i}$ (3)	$\frac{\sum_j Y_{ij}}{\sum (Y_i - K)}$ (4)	$\eta m_i \frac{\sum_j Y_{ij}}{\sum (Y_i - K)}$ (5)	$\frac{m_i^F}{r^F}$ (6)	$\eta m_i \left[\frac{\sum_j Y_{ij}}{\sum (Y_i - K)}\right] \frac{m_i^F}{r^F}$ (7)
34	6 8	0 253	6 06	0 413	64 5	0 266
35	7 0	0 164	3 83	0 268	69 1	0 185
36	7 2	0 198	4 49	0 323	73 7	0 238
36	7 2	0 238	5 38	0 387	78 3	0 303
37	7.4	0 054	1 19	0 089	82 9	0 074
38	7 6	1 505	32 23	2 449	87 6	2 145
39	7 8	0 229	4 78	0 373	96 8	0 361
40	8 0	1 599	32 55	2 604	101 4	2 640
41	8 2	0 043	0 85	0 070	106 0	0 074
41	8 2	0 236	4 68	0 384	110 6	0 425
42	8 4	1 371	26 57	2 232	115 2	2 571
43	8 6	0 100	1 88	0 162	124 4	0 202
44	8 8	0 705	13 05	1 148	129 0	1 481
44	8 8	0 055	1 03	0 091	133 6	0 122
46	9 2	0 048	0 85	0 078	142 9	0 111
46	9 2	0 403	7 13	0 656	147 5	0 968
48	9 6	0 022	0 37	0 036	156 7	0 056
48	9 6	0 003	0 05	0 005	161 3	0 008
49	9 8	0 250	4 15	0 407	165 9	0 675
50	10 0	0 013	0 21	0 021	175 1	0 037
51	10 2	0 145	2 32	0 237	179 7	0 426
51	10 2	0 015	0 24	0 024	184 3	0 044
51	10 2	0 001	0 01	0 001	188 9	0 002
51	10.2	0 097	1 55	0 158	193 5	0 306
53	10 6	0 077	1 20	0 127	207 4	0 263
55	11 0	0 051	0 75	0 083	221 2	0 184
56	11 2	0 375	5 45	0 610	230 4	1 405
Total		8 250 (dW/dT_P)		13 436 (dW/dT_D)		15 572 (dW/dT_G)

where $dm_1 = dm_2 = \quad = dm$ for a proportional tax Since the change in tax revenue, dT_p, is $\sum_n Y_i \, dm$, the marginal welfare cost is

$$dW/dT_p = \eta m_1 \frac{\sum_j Y_{1j}}{\sum_n Y_i} + \eta m_2 \frac{\sum_k Y_{2k}}{\sum_n Y_i} + \cdots \tag{13}$$

Thus, to calculate the marginal welfare cost, we multiply ηm_i for each rate class by the share of total wage and salary income accounted for by that class, and sum over all tax rate classes.

Columns 2 and 3 of table 2 present the results of these calculations for the proportional tax. The marginal welfare cost due to labor supply distortions is the sum of the figures in column 3, or about 8.3 percent of tax revenue The marginal administrative and compliance costs should be added to this figure. Average administrative and compliance costs are 2–2.5 percent of tax revenue, but presumably the marginal costs are less than this. Perhaps a figure of 0.7 percent is reasonable, and on that assumption the marginal welfare cost (due to labor supply distortions plus administrative and compliance costs) would be on the order of 9 percent of tax revenue The marginal cost of public funds would then be $1 09 per dollar of revenue.

To estimate the marginal welfare cost for a degressive tax that is added to the present system, it will be assumed that a certain amount of labor income is exempt from taxation Since actual taxable income under the federal income tax was only 61.4 percent of adjusted gross income in 1974, we shall assume that only 61 4 percent of total wage and salary income is subject to tax The formula for estimating the marginal welfare cost of labor supply distortions (which can be derived in a manner similar to equations [11]–[13] above) is then

$$dW/dT_D = \eta m_1 \frac{\sum_j Y_{1j}}{\sum_n (Y_i - K)} + \eta m_2 \frac{\sum_k Y_{2k}}{\sum_n (Y_i - K)} + \cdots, \quad (14)$$

where $\sum_n (Y_i - K)$ is total taxable wage and salary income ($455 billion), or 61 4 percent of total wage and salary income

Columns 4 and 5 give the estimates for the degressive tax The marginal welfare cost due to labor supply distortions is 13 4 percent of tax revenue. With marginal administrative and compliance costs included, the figure would be about 14 percent

For a graduated rate tax above an exemption, the formula when taxpayers are in different initial marginal rate classes is

$$dW/dT_G = \eta m_1 \frac{\sum_j Y_{1j}}{\sum_n (Y_i - K)} \frac{dm_1}{dr} + \eta m_2 \frac{\sum_k Y_{2k}}{\sum_n (Y_i - K)} \frac{dm_2}{dr} + \cdots, \quad (15)$$

where dr is the change in the average tax rate on total taxable wage and salary income (not the change in the average tax rate for each bracket). As mentioned earlier, to calculate the marginal welfare cost for such a tax, it is necessary to specify a particular change in the rate structure For ease of computation, we shall assume that the federal income tax rates are changed in a way that maintains constant share progressivity so that the dm_i/dr terms in (15) are equal to m_i^F/r^F, or each federal marginal tax rate divided by the average federal tax rate on taxable income[14]

Columns 6 and 7 give the estimates for this change in the tax structure The marginal welfare cost due to labor supply distortions is 15 6 percent, or slightly above 16 percent when marginal administrative and compliance costs are included Comparing the results of the three alternative methods of raising additional tax revenue, we can see that the marginal cost of public funds is greater the more progressive the change in the rate structure However, the exact magnitudes of all the estimates should be viewed as highly tentative since they are quite sensitive to the value of η assumed Unfortunately, empirical research has produced no consensus about the sizes of compensated labor supply elasticities, and my assumption of a value of 0 20 may be substantially off the mark

---

[14] Musgrave and Thin (1948) discuss share progressivity, although it is referred to as "liability progression" Constant share progressivity simply means that each person's share of total tax liability remains unchanged. The federal tax surcharge of 1968 had this characteristic.

As was probably clear in the last section, I consider the major use of an estimate of the marginal cost of public funds to be in the evaluation of government expenditures the marginal cost of public funds is the social opportunity cost of government spending This interpretation, however, immediately creates a difficult problem *which* estimate of marginal welfare cost should be used in the evaluation of a specific expenditure program? The difficulty arises because expenditure programs are not tied to particular tax changes in existing political institutions, and therefore the source of funds to finance any expenditure program can never really be known And matters become even more complex when still other changes in the income tax structure are considered, and when changes in nonincome taxes are also a source of potential revenue

It is important to recognize that it is literally impossible to determine the exact source of funds when governments use general fund financing (enacting tax and expenditure bills separately) In this type of situation what is clearly needed is a convention, or rule of thumb, that the applied economist can use in the evaluation of expenditure programs [15] It would appear that this convention should represent a judgment of the type of change in the tax structure that the government typically makes when more or less revenue is required Changes in the federal income tax over the past several years have tended to maintain the degree of progressivity (in the "constant share progressivity" sense), as in 1968, or to increase the degree of progressivity, as in 1964 and 1975 Thus, I would be prepared to argue that the assumption that public funds derive from progressive changes in the tax law (as in the third type of tax change considered above) would provide a reasonable convention to use in expenditure analysis Not everyone will agree with this judgment, but the need for *some* convention should be apparent

The importance of measuring tax distortions at the margin is not restricted solely to income taxes since other taxes are also potential sources of funds for the finance of expenditure programs There is reason to suspect, however, that marginal welfare costs are lower for income taxes than for other taxes Consider the case of selective excise taxes If we assume that these taxes do not influence labor supply, then the principal distortion of excise taxes is to influence relative output levels. The total welfare cost of this distortion can be estimated (ignoring second best problems, and assuming constant costs) by $\frac{1}{2}\varepsilon_d t^2 PQ$, where PQ is total expenditures on the taxed good, ε_d is the compensated demand elasticity, and t is the tax rate. The marginal welfare cost of an excise tax is then $\varepsilon_d t$ if it is assumed that quantity consumed is unresponsive to a change in the tax rate. Treating all goods subject to excise taxes as a composite good, we can use the estimates of Musgrave and Musgrave (1973, p. 459) of 0.8 for ε_d and 0.33 for t to estimate the marginal welfare cost. At 26 percent it is well above the marginal welfare cost of income taxation. If this is any indication of the magnitudes of marginal distortions of nonincome taxes, it suggests that income taxation may be the least expensive source of finance for expenditure programs.[16] Further research, however, is clearly needed to check this conjecture.

---

[15] Harberger (1971) has argued persuasively for the adoption of a convention in the analogous context of the choice of a social rate of discount.

Even though the marginal welfare cost of income taxes may be lower than for alternative sources of revenue, at 14–16 percent (for progressive changes in the tax structure), it is still a large part of the opportunity cost of government expenditures. Thus, it appears that the conditions under which expenditure policies can improve welfare are more restrictive than conventional analysis might suggest.

References

Aaron, Henry J. *Why Is Welfare So Hard to Reform?* Washington: Brookings Inst., 1973.

Bloch, Farrell. *The Allocation of Time to Market and Non-Market Work within a Family Unit.* Technical Report no. 114, Institute for Mathematical Studies in the Social Sciences, Stanford University, 1973.

Browning, Edgar K. "Labor Supply Distortions of Social Security." ***Southern Econ. J.*** 42 (October 1975): 243–52. (*a*)

——— "The Marginal Welfare Cost of Income Redistribution." Unpublished manuscript, University of Virginia, 1975. (*b*)

Buchanan, James M., and Tullock, Gordon. "Public and Private Interaction under Reciprocal Externality." In *The Public Economy of Urban Communities*, edited by J. Margolis. Washington: Resources for the Future, 1965.

Cain, Glen G., and Watts, Harold W. "Toward a Summary and Synthesis of the Evidence." In *Income Maintenance and Labor Supply*, edited by G. G. Cain and H. W. Watts. Chicago: Rand-McNally, 1973.

Council of Economic Advisers. *Economic Report of the President.* Washington: Government Printing Office, 1975.

Harberger, Arnold C. "Taxation, Resource Allocation, and Welfare." In *The Role of Direct and Indirect Taxes in the Federal Revenue System.* Princeton: Princeton Univ. Press for Nat. Bur. Econ. Res. and Brookings Inst., 1964.

---

[16] It also suggests, of course, that substituting income taxes for excise taxes while keeping total revenue unchanged would lower the total welfare cost of the tax system.

———. "On Measuring the Social Opportunity Cost of Public Funds." Arlington, Virginia. IDA Econ. Papers, 1971.

Johnson, David B , and Pauly, Mark V. "Excess Burden and the Voluntary Theory of Public Finance " *Economica* 36 (August 1969) · 269–76.

Musgrave, Richard A. "Cost-Benefit Analysis and The Theory of Public Finance " *J. Econ Literature* 7 (September 1969) 797–806.

Musgrave, Richard A , and Thin, Tun "Income Tax Progression, 1929–48 " *J.P.E.* 56 (December 1948) 498–514.

Musgrave, Richard A , and Musgrave, Peggy B *Public Finance In Theory and Practice* New York McGraw-Hill, 1973. ·

Pigou, A C *A Study in Public Finance*. 3d ed. London · Macmillan, 1947.

Vickrey, William S "General and Specific Financing of Urban Services " In *Public Expenditure Decisions in the Urban Community*, edited by H G. Schaller Washington Resources for the Future, 1963.

U.S. Bureau of the Census. *Current Population Reports* Series P-60, No 90. Washington. Government Printing Office, 1973

A Model of Optimal Human-Capital Accumulation and the Wages of Young High School Graduates

---

Charles Brown

Boston College

This paper estimates the parameters of Ben-Porath's model of optimal accumulation of human capital over the life cycle. A discrete-time version of the model is presented, and previous estimates of its parameters are considered. Using longitudinal data on wages of young white high school graduates, these parameters are reestimated. The point estimates of the two key parameters (the discount rate and the elasticity of investment costs with respect to investments) are implausible on a priori grounds, confirming the essentially negative conclusions of earlier studies.

In an oft-cited paper, Ben-Porath (1967) presented a formal model of optimal human-capital accumulation and showed that the optimal investment sequence implied an optimal profile for earnings. The purpose of this paper is to provide a more appropriate test of the model than is contained in Ben-Porath's (1970) subsequent paper, using data on white high school graduates from the Survey of Work Experience of Young Men (SWEYM).

In Section I, a discrete-time version of Ben-Porath's model and extensions proposed by Rosen (1973) are presented. In Section II, the underlying parameters of the model are considered in order to determine the plausible ranges for these parameters. In Section III, previous estimates of the model's parameters are considered. It is argued that deficiencies in these tests could produce the observed inconsistencies even

---

This paper is based on my thesis (Brown 1974). I am grateful to Richard Freeman, Zvi Griliches, James Medoff, and the referee for helpful suggestions. Financial assistance was provided by the National Institute of Education, U.S. Department of Health, Education, and Welfare, under research grant no. NE-G-00-0202. Responsibility for errors or omissions is, of course, my own.

[*Journal of Political Economy*, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved.

if the underlying model is correct. A brief discussion of the data in Section IV is followed by estimates of the model's parameters in Sections V and VI. The estimates in these two sections are quite similar; however, judged by the standards of Section II, they are questionable on a priori grounds. Conclusions are offered in Section VII.

I. A Model of Optimal Postschool Investment

It is assumed that the individual determines his human-capital investments in order to maximize the present value of earnings,

$$V = \sum_{t=0}^{N} Y(t)(1 + r)^{-t}, \qquad (1)$$

where t is the working age (i.e., length of working experience), $Y(t)$ is observed earnings at working age t, N is the working life, and r is the rate of interest, assumed constant.[1] Earning capacity at time t is equal to $R(1 + g)^t K(t)$, where $K(t)$ is the stock of human capital at time t, R is the rental price of human capital at time $t = 0$, and g is the rate of growth in the rental price of human capital, due to labor- (and human-capital-) augmenting technical progress, price inflation, etc. Again, R and g are assumed constant.[2] Moreover, the human capital is "disembodied"—that is, its return is independent of any working-hours choice by the individual.

Investment costs depend on the amount invested, $I(t) \equiv K(t + 1) - K(t)$, following Ben-Porath, it is assumed that

$$\text{cost of } I(t) = \frac{a}{\beta} I(t)^{\beta}(1 + g)^t,$$

a functional form which exhibits increasing marginal costs of investment[3] for $\beta > 1$.

This formulation of the cost of investment function makes two additional assumptions. The first is that the forces which generate g increase investment costs at the same rate as they increase the rental price of K. Since the costs of $I(t)$ are foregone output, if the time it takes to "learn" $I(t)$ is independent of the rate of technical progress, the costs of accumulating K will grow apace with its rental price. Convenience dictates that this assumption be maintained throughout.[4]

---

[1] Given the assumption of perfect capital markets, the Fisher "separation theorem" applies to human-capital investments, and eq. (1) is the proper maximand, independent of preferences about the time stream of consumption.

[2] Alternatively, g could be defined as the real rate of growth in the rental price of human capital and r as the real rate of interest.

[3] Increasing marginal costs are required to fulfill the second-order conditions for maximizing V.

A second neutrality assumption—*the* neutrality hypothesis in the literature—is also implicit in this formulation As K accumulates, the opportunity cost of a unit of time increases; hence, if the time it takes to "learn" $I(t)$ were independent of $K(t)$, the foregone-earnings cost of $I(t)$ would rise proportionally with $K(t)$ The model outlined above assumes instead that the *foregone-earnings* cost of $I(t)$ is independent of $K(t)$: one becomes more productive in acquiring human capital at the same rate that one becomes more efficient in producing output

By way of summary, we can rewrite equation (1) as

$$\begin{aligned} V &= \sum_{t=0}^{N} \left[RK(t) - \frac{a}{\beta} I(t)^{\beta} \right] [(1+g)/(1+r)]^{t} \\ &= \sum_{t=0}^{N} \left\{ R\left[K(0) + \sum_{s=0}^{t-1} I(s) \right] - \frac{a}{\beta} I(t)^{\beta} \right\} D^{-t}, \end{aligned} \tag{2}$$

where $K(0)$ is the human-capital stock on entry into the labor market, and $D = (1+r)/(1+g) \simeq 1 + r - g$

Maximizing V with respect to $I(0), I(1), \quad , I(N)$ requires that[5]

$$R \sum_{j=t+1}^{N} D^{-j} - aI(t)^{(\beta-1)} D^{-t} = 0 \rightarrow R \sum_{j=1}^{N-t} D^{-j} - aI(t)^{(\beta-1)} = 0$$

Since the first term is the present value of the earnings stream generated by an additional unit of K, while the second term is the marginal cost of acquiring an additional unit of K, this equation is simply the usual requirement that marginal return equal marginal cost For given t, the marginal return is constant, while marginal cost is a function of $I(t)$ The time path of investments is determined by the horizontal marginal-return function sliding downward and intersecting the stationary marginal-cost function at lower levels of investment (see fig 1)

---

[4] Becker (1964, p 53) interprets "neutral" technological progress differently. returns from all investments grow at rate g, but costs are unchanged He argues that this would explain the growth in human-capital investments over time Whatever the merit of his argument in explaining rising educational levels, its appeal in relation to *on-the-job* investments depends on these investments having grown over time, a more questionable proposition

[5]

$$\frac{\partial V}{\partial I(t)} = \sum_{j=0}^{N} R \frac{\partial[\sum_{s=0}^{j-1} I(s)]}{\partial I(t)} D^{-j} - aI(t)^{(\beta-1)} D^{-t} = 0$$

But

$$\frac{\partial[\sum_{s=0}^{t-1} I(s)]}{\partial I(t)} = \begin{cases} 1 \text{ for } j \geq t+1 \\ 0 \text{ for } j \leq t \end{cases}.$$

The equation in the text follows directly

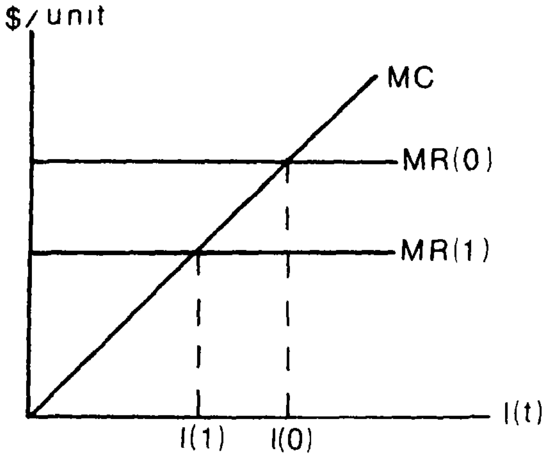


FIG 1 —Determination of optimal investments

Solving for $I(t)$ yields

$$I(t) = \left(\frac{R}{a}\sum_{j=1}^{N-t} D^{-j}\right)^{[1/(\beta-1)]} \equiv \left(\frac{R}{a}P(N-t)\right)^{\alpha} \tag{3}$$

Here $P(s)$ is the present value of a dollar per year for s years, and $\alpha \equiv 1/(\beta - 1)$

Unfortunately, $I(t)$ is not directly observable However, given $K(0)$, a series of $I(t)$ will imply a path for $Y(t)$

$$Y(t) = (1+g)^t\left\{RK(0) + C\left[\sum_{s=0}^{t-1} P(N-s)^{\alpha} - \frac{1}{\beta}P(N-t)^{\beta\alpha}\right]\right\}, \tag{4}$$

where $C \equiv R(R/a)^{\alpha} = a(R/a)^{\alpha\beta}$ Rather than focusing on the level of Y (as in equation [4]), we can also find the increase in Y implied by the optimal investment series

$$\begin{aligned}\Delta Y(t) &\equiv Y(t) - (1+g)Y(t-1)\\ &= (1+g)^tC\left\{P(N-t+1)^{\alpha} - \frac{1}{\beta}\right.\\ &\qquad\left.[P(N-t)^{\alpha\beta} - P(N-t+1)^{\alpha\beta}]\right\}\end{aligned} \tag{5}$$

Thus, both the level and the change in Y are functionally related to one observable variable (working age) and four parameters (β, D, C, and g) which are assumed to be the same for all individuals[6]

---

[6] Of course, the sample could be divided into subgroups and the hypothesis tested that some or all of the parameters were equal across groups (see Rosen 1973) Dividing the sample into two groups, according to IQ scores or according to SES, produced no statistically significant differences in parameter estimates between groups (see Brown 1974)

The model outlined above describes the optimal age-earnings profile of an individual over time; it is less adequate as a description of the earnings of individuals from different cohorts over time, which the SWEYM provides Equation (2) makes the worker's earnings a function of his work age t, but ignores the effect of cohort Thus, the model predicts that if father and son are otherwise identical, their earnings at $t = 0$ (or any other t) are equal However, the effects of labor-augmenting technical progress, secular price inflation, and differences in the initial stock $K(0)$ which each cohort brings to the labor market will make earnings of identical individuals in different cohorts diverge Let v (for vintage) be the index for cohorts The most convenient cohort index defines $v =$ – (no periods since the cohort's entry into the labor market)

Rosen incorporates the influence of cohort with two modifications of the basic model First, the initial capital stock $K(0)$ is allowed to be a function of the individual's cohort The $K(0, v)$ might increase over v, for example, if skills previously learned in the labor market came to be learned in school, or if workers learn additional skills while in school via part-time employment[7] On the other hand, in a sample selected according to years of schooling, $K(0, v)$ might decline with v as average schooling levels increased throughout the economy, the "ability" range from which the typical high school graduate is drawn may decline Welch (1973) found evidence of positive vintage effects, suggesting that the first set of influences may predominate A second type of vintage effect is included by assuming that both the rate of return and the cost of investment at $t = 0$ grow at rate g

$$R(v) = R(0)(1 + g)^v,$$
$$a(v) = a(0)(1 + g)^v$$

This means that for the vth cohort, $(1 + g)^v$ can be factored out of the expression for earnings and thus from the present value expression (eq [2]) The optimal investment path is unchanged since, at any time t, $(1 + g)^v$ multiplies both the marginal-return term and the marginal-cost term Therefore, the earnings for an individual of working age t will be

$$Y(t, v) = (1 + g)^{v+t} \left\{ RK(0, v) + C(0) \left[\sum_{s=0}^{t-1} P(N - s)^{a} - \frac{1}{\beta} P(N - t)^{a\beta} \right] \right\} \tag{4'}$$

For individual cross sections, $v + t$ is constant, if cohorts are indexed appropriately, $v + t = 0$, and (4′) reduces to (4) When the available data are a series of cross sections, $v + t$ is a constant for each cross section

---

[7] It is worth emphasizing that changes which increase the ability to learn skills in the labor market (i.e., reduce a) are not examples of this type

When only a single cross section of Y is available, we can derive an expression for Y-gains only on the assumption that $K(0; v)$ is *constant across cohorts*[8]

$$\begin{aligned}\Delta^* Y(t, v) &\equiv Y(t, v) - Y(t - 1, v + 1)\\ &= (1 + g)^{v+t} C(0) \qquad (5^*)\\ &\quad \times \left\{P(N - t + 1)^{\alpha} - \frac{1}{\beta}\left[P(N - t)^{\alpha\beta} - P(N - t + 1)^{\alpha\beta}\right]\right\}\end{aligned}$$

If a series of cross sections is available, we can difference over time *for each individual* to get

$$\begin{aligned}\Delta Y(t, v) &\equiv Y(t, v) - (1 + g) Y(t - 1, v)\\ &= (1 + g)^{v+t} C(0) \qquad (5')\\ &\quad \times \left\{P(N - t + 1)^{\alpha} - \frac{1}{\beta}\left[P(N - t)^{\alpha\beta} - P(N - t + 1)^{\alpha\beta}\right]\right\}.\end{aligned}$$

II. Determining Plausible Ranges for the Model's Parameters

Before considering estimates of the parameters of the model, it is useful to ask whether there are limits within which each of these parameters should lie

Since g is the growth rate of the rental price of human capital in nominal terms, it includes the influence of both inflation and changes in the real rental price If the stock of human capital per worker were constant over time, growth in a suitable per capita earnings index could plausibly be taken as a measure of g[9] If we are willing to assume that human capital per worker is not declining over time,[10] g is less than or equal to the growth in average earnings On the other hand, unless the rental price of human capital is declining in real terms, g will be greater than or equal to the rate of price inflation

Since $D \simeq 1 + r - g$, finding a plausible range for D depends on finding bounds for r If r is taken to be a lending rate rather than a borrowing rate, and does not include a "risk premium," then r should roughly equal g.[11] If, on the other hand, it represents a borrowing rate or includes a risk premium, somewhat higher values become plausible. Both Ben-Porath (1970, p. 141) and Thurow (1970*b*, p. 74) have argued that, for human capital investments, r may plausibly be 15 percent or more.

---

[8] Differencing across cohorts makes sense only if our observations are group-means rather than individual-level data

[9] A change in the average human-capital stock per worker would occur if either the age structure of the work force or age-specific human-capital stocks change. Vintage-related differences in $K(0)$ would be an example of the latter change

[10] This assumption is more plausible in the sample period being studied (1966–69) than it might be in other (e g , more recent) periods.

Theory requires β to be greater than 1.0 if investments are to be positive and finite. While any value of β greater than 1.0 is theoretically admissible, the optimal-investment equation (2) gives some insight into plausible values for β. The elasticity of investment with respect to R/a is $\alpha \equiv 1/(\beta - 1)$. Suppose that a is reduced by 10 percent—by government subsidy for on-the-job training, for example—and R could be maintained. How much would each $I(t)$ increase? If $\beta = 2$, $I(t)$ increases by 10 percent for all t, while if $\beta = 1.10$, $I(t)$ doubles. Most readers would regard the latter figure as implausible. However, it is hard to be sure that this reflects the implausibility of β being as low as 1.10, or skepticism that R could remain roughly constant if a reduction in a were successfully engineered.[12]

Finally, it does not appear possible to place any bounds on C other than the obvious fact that it must be positive.

III. Estimates from Previous Research

The first estimate of the parameters of a model like this was reported by Ben-Porath (1970). While exogenous growth in the rental price of human capital is not discussed explicitly, the omission is of little consequence since the data are cross-sectional, any estimate of g would be embedded in (and not identifiable from) the estimate of D (see eq. [5*]).

Rather than confront wage or earnings data directly, Ben-Porath examined the ability of the model to predict investment costs [$= (a/\beta)\, I(t)^{\beta}$]. His investment-cost data were Mincer's estimates of 1958 training costs for males by age and level of education (see Mincer 1962). He found that the model generated estimates of d in the low 20s and β very near 1.0 for the high school graduates. However, investments decline more rapidly than the model predicted.

---

[11] Feldstein (1973, p. 21) argues that the lending rate available to most investors is not very different from the rate of growth of nominal output per man.

[12] An analogy with the neoclassical model of physical investment may be helpful here. The voluminous discussion of Jorgenson's (1963) investment equation focuses precisely on the *partial* elasticity of desired capital stock with respect to user cost, if stimulating investment leads to lower output prices, this partial elasticity may overstate the total response of the capital stock to reductions in user cost. What is "missing" is an equation relating competitive output price to the user cost of capital. Similarly, if demand for human-capital services is inelastic, any increase in I caused by a reduction in a would be "choked off" by declining R.

Ben-Porath's paper seems to have discouraged further tests of his model Even research whose theoretical foundation is based on Ben-Porath's model has avoided using the implied functional form in empirical tests (e.g , Lillard 1973, and, less clearly, Haley 1973)

However, the negative results he reports may well reflect flaws in the testing procedure rather than inadequacies of the model The overrapid decline in investment costs may be illusory rather than real

Since the investment-cost series is based on *cross-sectional earnings* data, either vintage effects in $K(0)$ or variation in work year could produce the illusion of declining investments The danger of vintage effects was recognized by Ben-Porath himself (1970, p 141) if $K(0)$ is smaller for older members of the cross section, and this fact is ignored, a decline in earnings gains in the cross section will be taken to reflect declining investments A similar danger, due to variations in work year, was pointed out by Thurow (1970*a*, p 154) if work years increase at young ages and level off thereafter, earnings gains will level off even if investments remain constant, Mincer's procedure, however, would take the decline in earnings gains as reflecting declining investments

Rosen (1973) suggested an additional difficulty with Ben-Porath's test of his model He noted that any attempt to infer human-capital investments from observed earnings will be conditional on the assumed cost-of-investment function Mincer's estimates assume that the human-capital investment function exhibits constant average costs But, Rosen argued, this assumption is inconsistent with Ben-Porath's model of optimal investment. it would mean that marginal costs as well as marginal returns are constant, and optimal investment would be either infinite, zero, or indeterminate Abandoning the investment-cost series, Rosen attempted to test the model directly on cross-sectional earnings data from the 1960 census However, he imposes the restriction $\beta = 2$ on the model. The resulting point estimates of d are *negative* for white high school graduates

An appropriate test of the model therefore requires a *repeated* cross section of data on *wages*, so that the pitfalls of vintage effects in $K(0)$ and changes in work year, and the conceptual problems Rosen pointed out, can be avoided The SWEYM provides such data, moreover, by studying young men, we can observe the period in the life cycle when most post-schooling investments are made

IV. The Data

Data for estimating equations (4′) and (5′) are available from the SWEYM [13] A 5,225-member sample of the civilian noninstitutional male population age 14–24 was interviewed originally in 1966; reinterviews were conducted in each of the 3 following years. From these 5,225 potential observations, the sample analyzed here was selected. It included white respondents who (1) had completed exactly 12 years of schooling in 1966, (2) were out of school at each survey date, and (3) reported a wage at each survey date. The resulting sample size was 323.[14] Obviously, any selection procedure reflects judgments on the importance of sample homogeneity versus sample size; the present sample clearly reflects an emphasis on homogeneity.

---

[13] For a more detailed description of the data, see Parnes, Miljus, and Spitz (1970, chap. 1 and appendix A)

V. Explaining Wage Levels

Equation (5′) was estimated from the data in this sample, with $Y(t-1, v) = 1966$ wage, and $Y(t, v) = 1969$ wage. The work life N was chosen so that a retirement age of 65 was assumed. Rather than estimating g directly, alternative values were imposed, and B, D, and c estimated conditional upon g. This decision involves two conceptually separate issues—the justification for imposing a value of g rather than estimating it directly, and the choice of the appropriate value for g.

In principle, g could be estimated directly by modifying equation (5′). If $Y(t)$ replaces $\Delta Y(t)$ as the dependent variable, and $Y(t-1)$ is included as an independent variable, the coefficient of $Y(t-1)$ would be an estimate of $(1+g)$. Unfortunately, Y is subject to measurement error and "transitory" influences (which are formally equivalent to measurement errors), so that estimates of the coefficient of $Y(t-1)$ are biased toward zero. In fact, this bias drives the estimated coefficient below 1.0, so that the implied value of g is negative.[15]

Although our primary interest is not in g, biased estimates of the coefficient of $Y(t-1)$ result in biased estimates of C for given β and D (and, presumably, error in the choice of β and D as well). Imposing a value of g can provide better estimates of other parameters.

Suppose the "true" model is

$$Y(t) = CX + (1+g)Y(t-1) + \varepsilon,$$

where

$$X = P(N-t+1)^{\alpha} - \frac{1}{\beta}\left[P(N-t)^{\alpha\beta} - P(N-t+1)^{\alpha\beta}\right]$$

Since X is a decreasing function of t, and $Y(t-1)$ grows with t, X and $Y(t-1)$ are correlated. If we estimate

$$Y(t) - (1+\tilde{g})Y(t-1) = \hat{C}X + \epsilon \tag{6}$$

(i.e., impose $g = \tilde{g}$ and estimate C), the bias in estimating C will be a monotonic function of $g - \tilde{g}$. Note, however, that estimating g and C

---

[14] Wages less than \$1.00 or greater than \$7.50 were excluded as implausible.

[15] Of course, estimates of g less than 1.0 are consistent with other interpretations (e.g., violation of the "neutrality" hypothesis).

directly by least squares gives the same estimate of C as imposing $\tilde{g} = \hat{g}$ (where $\hat{g}$ is the least-squares estimate of g). Therefore, if $\hat{g}$ is a biased estimate of g, and there is a value $\tilde{g}$ such that $\hat{g} < \tilde{g} \leq g$, imposing $\tilde{g}$ and estimating (6) gives a better estimate of C than estimating g and C directly[16]

The choice of the correct value of g is complicated by the fact that there are two different growth rates of the rental price of human capital appearing in equation (5′)—one appearing explicitly on the left-hand side, the other embedded in D The g on the left-hand side is simply a price inflator (reflecting changes in the rental price of human capital), and the actual value of g between 1966 and 1969 is the correct choice The g on the right-hand side of the equation reflects the rate at which the rental price is expected to grow over the remaining horizon, so a longer-run value of g is appropriate In order to estimate equation (5′), only the first g must be chosen in advance, the latter is needed only to identify r for given D

Three values were chosen for the rate of growth in the rental price of human capital between 1966 and 1969. 0.13, 0 19, and 0 32 The first represents the rate of growth of (consumer) prices during the period, while the second is the rate of growth in average gross hourly earnings in the private economy The third value is based on the ratio of wages of sample members age 20–24 in 1969 to wages of sample members age 20–24 in 1966 The discussion in Section II suggests that these choices span the range of reasonable values for g

Equation (5′) was estimated by searching over various values of β and D, and selecting the β, D pair which minimized the sum of squared residuals A grid of values of D ranging from 1 01 to 1 50 was established Since the "period" of analysis is 3 years, these correspond to annual discount rates from 0 003 to 0 15 For given D, a golden-section search[17] of values of β from 1 03 to 2 00 was undertaken to find the optimal point estimate of β Fortunately, there was no problem of multiple local minima

The results of the estimation are summarized in table 1 Three conclusions emerge

1 While the point estimates of β depended on the value of g imposed, the point estimate of D which minimized the sum of squared residuals was

---

[16] Formally, $Y(t-1)$ is an omitted variable whenever $\hat{g} \neq g$, and the bias in the estimate of C is the product of the coefficient of X regressed on $Y(t-1)$ and the coefficient of $Y(t-1)$ in the "true" equation $Y(t) - (1+\hat{g})Y(t-1) = CX + (g-\hat{g})Y(t-1) + \epsilon$ Since that coefficient is equal to $(g-\hat{g})$, a better imposed value reduces the bias of estimating C The situation differs from the usual problem of prior knowledge about coefficients (Theil 1971, pp 346–47), in that the least-squares estimate of C is not unbiased.

[17] A golden-section search is an efficient technique for finding a local minimum for a function (see Jacoby, Kowalik, and Pizzo 1972, chap 3).

TABLE 1

POINT ESTIMATES OF β AND D IN EQUATION (5′)

	$g = 0.13$ (CPI)	$g = 0.19$ (AGHE)	$g = 0.32$ (Internal)
Equation (5′):			
Estimate of β	1.792	1.468	1.159
Estimate of D	1.010	1.010	1.010
Sum of squared residuals	341.69	357.01	389.24
Quadratic function of age			
Sum of squared residuals	341.58	356.90	389.18

1.01 for all three values of D. While the plausibility of the point estimates of β appears sensitive to choice of g, the point estimates of D are disarmingly low.[18]

2. The data do not allow us to estimate β and D very precisely. Since β and D enter equation (5′) nonlinearly, a standard procedure for finding the confidence region is that based on the likelihood ratio. Choosing values of β and D other than the "optimal" point estimates can be thought of as imposing two restrictions on the estimation of the equation. Let SSR_{Ω} be the sum of squared residuals at the optimal point estimates, and SSR_w be the sum of squares when $\beta = \beta^*$, $D = D^*$. Then the likelihood ratio is equal to

$$\left(\frac{SSR_w}{SSR_{\Omega}}\right)^{-(n/2)},$$

where n is the number of observations. Since we have imposed two restrictions, the quantity

$$n \log \frac{SSR_w}{SSR_{\Omega}}$$

is distributed approximately as $\chi^2(2)$ (Theil 1971, pp. 396–97). To find the β^*, D^* pairs which lie within a 90 percent confidence region $(\chi^2_{0.90}(2) = 4.61)$, we solve the following inequality for SSR_w:

$$n \log \frac{SSR_w}{SSR_{\Omega}} \leq 4.61$$

$$SSR_w \leq \left(1 + \frac{4.61}{n}\right) SSR_{\Omega} \text{ (approx.)}$$

---

[18] The 1966 wage variable was created without direct knowledge of "usual hours on present job"—a variable which was available in the 1967–69 surveys. Usual hours were constructed as "hours last week" − "overtime" + "unusual reduction" in workweek. To be certain that this potential noncomparability did not distort the results, eq. (5′) was reestimated using only 1967–69 wages. Point estimates of β were about 0.20 higher, but the point estimates of D remained at 1.01.

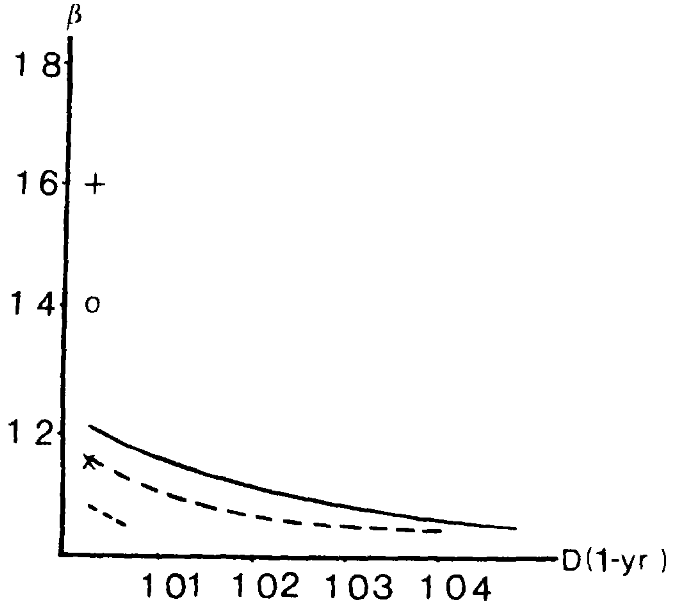


FIG. 2.—90 percent confidence regions for β and D in eq. (5'). *Solid line* = CPI (g = 0.13), *long dashes* = AGHE (g = 0.19), *short dashes* = Internal (g = 0.32).

In figure 2, approximate 90 percent confidence regions for β and D are shown for each of the three values of g. Only β, D pairs in which both parameters are near 1.0 lie outside these regions. Thus, while the point estimates are unsatisfactory, the confidence regions include almost any plausible β, D pair.

3. Equation (5') fits the data roughly as well as an equation in which wage gains are a simple quadratic function of age. The rough comparability of the two forms is not surprising, given the limited age range covered by the data. Equation (5'), however, flattens out without turning down as $t \rightarrow N$. This property is consistent with existing data (see Mincer 1974, p. 70) and is difficult to impose on the quadratic alternative.

Arguably, the sample on which these conclusions is based is inappropriate, in that the large wage gains observed among the youngest respondents may reflect the process of finding jobs consistent with their skills as well as the augmentation of skills which the model emphasizes. Ornstein (1971, p. 417) has concluded that such adjustment "takes place very quickly, mostly within a year of entry [into the labor force] and almost all of it within two years after entry."

To explore this possibility, equation (5') was estimated from a sample which included only those respondents who had begun their first job after high school at least 2 years before the 1966 survey. However, this experiment produced parameter estimates which were, if anything, even less satisfactory. the estimates of D remained at 1.01, while point estimates of β were lower than in the larger sample.

VI. Explaining Wage Levels

Focusing on changes in wages or earnings is a convenient way of dealing with differences in the initial capital stock, $K(0)$ With the data available from the SWEYM, a real price is paid for this convenience, since data for 1967 and 1968 were effectively discarded. It is possible to make use of the data reported in these 2 years by estimating equation (4′), which describes the time path of the level of wages implied by an optimal investment program

As before, when the rate of growth of the rental price of human capital is not constant, $(1 + g)^{v+t}$ is replaced by an index of the actual growth rate Wages were deflated by this index (for each of our three measures of g), and the "constant dollar" wage was used as dependent variable.

For given β and D, the term in brackets in (4′) can be constructed for each individual in each year, since it is a known function of the individual's age If we were willing to assume that there are no vintage effects in $K(0)$, an intercept could capture $RK(0)$ We can allow $K(0)$ to differ across cohorts by adding a set of cohort-specific dummy variables

Given an additive error term with the usual Gauss-Markov properties,[19] equation (4′) can be estimated directly by ordinary least squares For each β, D pair, we can compute the cohort-specific initial-stock effects, the coefficient C, and the sum of squared residuals, after searching over β and D,[20] we can select the value which minimizes the sum of squared residuals

The values of β and D which minimized the sum of squared residuals are presented in table 2 The values of D are *1-year* discount factors. Comparing these results with those in table 1 reveals that the two sets of point estimates are strikingly similar

If the residuals from equation (4′) satisfy the usual conditions, the β, D pair which minimizes the sum of squared residuals will be maximum-likelihood estimates of the two parameters While it is reasonable to assume that the disturbances are independent across individuals, it is questionable whether the disturbances for any individual are independent over time. If we "stack" the observations so that the four observations for any individual appear sequentially, our dependent variable Y is a 1,292-element (= 4 years × 323 individuals) vector The matrix of independent variables has nine columns (= eight cohort dummies[21] and the nonlinear function of β and D in eq [4']) and 1,292 rows, while the coefficient vector γ has nine elements (eight vintage effects and the estimate of C) Letting i index individuals and τ index time, we have

---

[19] This is not an innocent assumption, since it implies that the error for an individual at working age t is independent of that at $t - 1$. This assumption is examined below

[20] The search was over a rectangular grid of β, D values. The golden-section search over β was abandoned due to cost

$$\begin{bmatrix} Y_{1,1} \\ Y_{1,2} \\ Y_{1,3} \\ Y_{1,4} \\ Y_{2,1} \\ \\ Y_{323,4} \end{bmatrix} = \begin{bmatrix} \\ X_{1,1,1} & & X_{1,1,9} \\ \\ X_{323,4,1} & & X_{323,4,9} \\ \end{bmatrix} \begin{bmatrix} \\ \gamma_1 \\ \\ \gamma_9 \end{bmatrix} + \begin{bmatrix} \\ u_{1,1} \\ \\ u_{323,4} \\ \end{bmatrix}$$

or $Y = X\gamma + U$ Of course, individuals can be grouped by cohort to provide further structure to the X matrix, but that is not essential here

Define $\sum^* = E(UU')$, $\sum^*$ is a symmetric matrix with 1,292 rows and columns Independence of disturbances across individuals means that $\sum^*$ can be written as a block-diagonal matrix, with the 4 × 4 submatrix $\sum_i$, $i = 1,$, 323 as the ith block If we assume that $\sum_i$ is identical across i, and let $\sum$ be this common submatrix, we have

$$\Sigma^* = \begin{bmatrix} \Sigma & & & 0 \\ & \Sigma & & \\ & & \Sigma & \\ 0 & & & \Sigma \end{bmatrix}$$

The sums of squared residuals can be used to select maximum-likelihood estimates of β and D only if $\sum = \sigma^2 I$

A more general procedure is to find a maximum-likelihood estimate of $\sum$ and γ conditional on the various β, D pairs and ask whether $\sum = \sigma^2 I$ appears to be a reasonable restriction on $\sum$

The computational procedure is conceptually straightforward for given β and D, estimate γ by ordinary least squares Estimate $\sum$ from the least-squares residuals $e_{1,1},$, $e_{1,4}, e_{2,1},$, $e_{323,4}$ as

$$\hat{\Sigma}_{j,k} = \frac{1}{323} \sum_{i=1}^{323} e_{i,j} e_{i,k}$$

One can use this $\hat{\sum}$ as the basis for a second round of weighted least squares and obtain a revised coefficient vector The new coefficient vector can be used to obtain another estimate of the variance-covariance matrix, and this in turn used in another round of weighted least squares for still another coefficient vector. If this iterative procedure converges, the $\hat{\gamma}$ and $\hat{\Sigma}$ are maximum-likelihood estimates of γ, Σ. Happily, in the present sample, the iterative process produced almost no revision in $\hat{\gamma}$ or $\hat{\Sigma}$ after the first weighted regression.

---

[21] There are eight cohorts in the sample, since it contains individuals age 17–24 in 1966.

The estimated variance-covariance matrices had substantial (and nearly equal) off-diagonal elements.[22] For example, with $g = 0.19$, $\beta = 1.40$, and $D = 1.003$, the estimated variance-covariance matrix was

$$\begin{matrix} 0.644 & 0.373 & 0.359 & 0.330 \\ 0.373 & 0.564 & 0.410 & 0.384 \\ 0.359 & 0.410 & 0.636 & 0.514 \\ 0.330 & 0.384 & 0.514 & 0.795 \end{matrix}$$

We can test the restriction $\Sigma = \sigma^2 I$ as follows. If the restriction is true, ordinary least squares provides the maximum-likelihood estimate for the coefficient vector, and the estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{nT} \sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2$$

Let e_i be defined as the T-vector whose elements are $e_{i,1}, \ldots, e_{i,T}$ (i.e., the vector of residuals for the ith individual), and k be the number of columns in X ($= 9$), then the log-likelihood function is

$$\begin{aligned} L_0 &= -\frac{nk}{2} \log 2\pi - \frac{n}{2} \log |\hat{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} e_i' \hat{\Sigma}^{-1} e_i \\ &= -\frac{kn}{2} \log 2\pi - \frac{nT}{2} \log \hat{\sigma}^2 - \tfrac{1}{2} nT \end{aligned} \tag{7}$$

Given an unrestricted matrix Σ, the iterative procedure described above is appropriate. Then,

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} e_i e_i',$$

where e_i is now the T-vector of residuals using the final (maximum-likelihood) estimate of the coefficient vector. The log-likelihood function is

$$\begin{aligned} L_1 &= -\frac{kn}{2} \log 2\pi - \frac{n}{2} \log |\hat{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} e_i' \Sigma^{-1} e_i \\ &= -\frac{kn}{2} \log 2\pi - \frac{n}{2} \log |\hat{\Sigma}| - \tfrac{1}{2} nT \end{aligned} \tag{8}$$

---

[22] The near equality of off-diagonal elements suggests that U_{it} may be of the variance-components form

$$U_{it} = W_i + V_{it}$$
$$EW_i = EV_{it} = EW_i V_{it} = 0.$$

The implied Σ is then $EU_i U_i' = \operatorname{Var}(W_i)\, jj' + \operatorname{Var}(V_{it}) I$ where j is a T-vector of ones.

TABLE 2

POINT ESTIMATES OF β AND D IN EQUATION (4′)

	CPI	AGHE	Internal
Equation (4′) (OLS)			
Estimate of β	1 600	1 400	1 150
Estimate of D	1 003	1 003	1 003
Sum of squared residuals	901 61	852 38	760 08
Test statistic for null hypothesis $\Sigma = \sigma^2 I$	640	653	632
Equation (4) (GLS)			
Estimate of β	1 600	1 400	1 150
Estimate of D	1 003	1 003	1 003
$\lvert\hat{\Sigma}\rvert$	0 03268	0 026506	0 016941

Then the logarithm of the likelihood ratio is

$$\lambda = \frac{n}{2}\left(\log|\hat{\Sigma}| - T\log\hat{\sigma}^2\right)$$

The test statistic -2λ reported in table 2 is distributed (asymptotically) as $\chi^2(9)$ Therefore, we can reject the hypothesis $\Sigma = \sigma^2 I$ with a high degree of confidence

The results of estimating equation (4′) with Σ unconstrained, for a grid of β, D pairs, are also presented in table 2 The maximum-likelihood estimates of β, D are the β, D pair which minimize $|\hat{\Sigma}|$ (see eq [8]) The point estimates labeled "GLS" in table 2 are these maximum-likelihood estimates Comparison of the OLS and GLS entries reveals that the point estimates of β and D are virtually unchanged by the use of the generalized least-squares procedure

Table 2 can be used to construct a confidence region for β and D, using the likelihood ratio For any β and D, the log-likelihood equals $n/2\ |\hat{\Sigma}|$ plus a constant which does not depend on β and D If we let $\hat{\Sigma}_0$ be the $\hat{\Sigma}$ matrix at the optimal β and D, and $\hat{\Sigma}_1$ be the $\hat{\Sigma}$ matrix for some other β, D pair, we have

$$\lambda = \frac{n}{2}\log\left(\frac{|\hat{\Sigma}|_0}{|\hat{\Sigma}|_1}\right),$$

and -2λ is distributed asymptotically as $\chi^2(2)$ Therefore, for a given confidence level, we can solve for the value of $\log|\hat{\Sigma}|$, and hence for the value of $|\hat{\Sigma}|$, which is on the boundary of the confidence region.

Figure 3 presents 90 percent confidence regions As in figure 2, it is clear that only β, D pairs in which both elements are near 1 0 lie outside these regions

Equation (4′) was also estimated for the subset of respondents who had

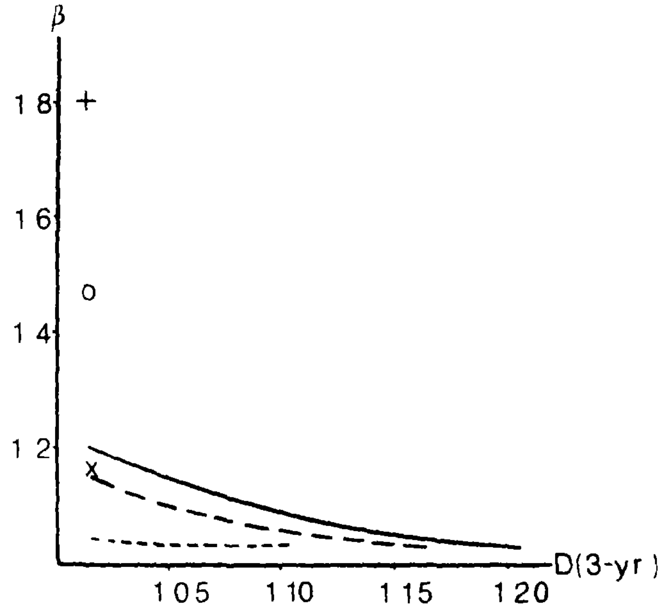


FIG 3—90 percent confidence regions for β and D in eq (4′) *Solid line* = CPI (g = 0 13), *long dashes* = AGHE (g = 0 19), *short dashes* = Internal (g = 0 32)

begun their first job after high school at least 2 years before the 1966 survey Once again, point estimates of D were unaffected, but estimates of β were somewhat lower in this sample

VII. Conclusions

The SWEYM data on white high school graduates allow a more appropriate test of the optimal investment model than is possible with cross-sectional data While the point estimates presented here differ from those reported by Ben-Porath (1970), the general performance of the model is equally discouraging The SWEYM estimates of β were higher than Ben-Porath's, though somewhat sensitive to choice of g Estimates of D, however, were consistently near 1 0 Unless one is prepared to accept the idea of interest rates for human-capital investments near long-run g, these point estimates of D must be regarded as implausibly low Confidence regions about these point estimates were uniformly large, reflecting both the importance of transitory influences in the data and the peculiarities of the nonlinear estimating equations which the model generates

In commenting on the findings of Ben-Porath's paper, Mincer (1970) emphasized the restrictiveness of the neutrality hypothesis as a likely villain[23] Unfortunately, unless the neutrality hypothesis is maintained, there is no closed-form solution for the optimal sequence of investments and, hence, no explicit optimal earnings profile We are left with a specific version of the model which generates questionable parameter estimates and a more general model which is difficult to test directly

References

Becker, Gary *Human Capital* New York Nat Bur Econ Res , 1964.

Ben-Porath, Yoram "The Production of Human Capital and the Life Cycle of Earnings " *J P E* 75, no 4, pt 1 (August 1967) 352–65

——— "The Production of Human Capital over Time " In *Education, Income, and Human Capital*, edited by W Lee Hansen New York Nat Bur Econ Res , 1970

Brown, Charles "Young Men Human Capital and Earnings " Ph D dissertation, Harvard Univ , 1974

Feldstein, Martin "Social Security, Induced Retirement, and Aggregate Capital Accumulation " Discussion Paper no 312, Harvard Inst Econ Res , 1973

Haley, William J "Human Capital The Choice between Investment and Income " *A E R* 63 (December 1973) 929–40

Jacoby, S L. S , Kowalik, J S , and Pizzo, J T *Iterative Methods for Nonlinear Optimization Problems* Englewood Cliffs, N J Prentice-Hall, 1972

Jorgenson, Dale W "Capital Theory and Investment Behavior " *A E R* 53 (May 1963) 247–59.

Lillard, L A. "Human Capital Life Cycle of Earnings Models A Specific Solution and Estimation " Working Paper no 4, Nat Bur Econ. Res , 1973

Mincer, Jacob "On-the-Job Training Costs, Returns, and Some Implications." *J.P E* 70, no 5, pt 2 (October 1962) 50–79

——— "Comment" on paper by Y Ben-Porath In *Education, Income, and Human Capital*, edited by W Lee Hansen New York Nat Bur Econ Res , 1970.

——— *Schooling, Experience, and Earnings* New York Nat Bur Econ Res , 1974.

Ornstein, Michael *Entry into the American Labor Force* Baltimore Center Soc Org Schools, John Hopkins Univ , 1971

Parnes, Herbert S , Miljus, Robert C , and Spitz, Ruth S *Career Threshholds.* Manpower Research Monograph no 16, vol 1 Washington Government Printing Office, 1970.

Rosen, Sherwin "Income Generating Functions and Capital Accumulation." Discussion Paper no 306, Harvard Inst Econ Res , 1973

Theil, Henri. *Principles of Econometrics* New York Wiley, 1971

Thurow, Lester "Comment" on paper by Y Ben-Porath In *Education, Income, and Human Capital*, edited by W Lee Hansen. New York Nat Bur Econ. Res , 1970. (*a*)

——— *Investment in Human Capital* Belmont, Calif. Wadsworth, 1970 (*b*)

Welch, Finis. "Black-White Differences in Returns to Schooling " *A.E R.* 63 (December 1973) 893–907.

[23] Some tentative results in Brown (1974) suggest that the problem may lie deeper than simply relaxing the neutrality hypothesis

Determining Legislative Preferences on the Minimum Wage: An Economic Approach

---

Jonathan I Silberman and Garey C. Durden

Old Dominion University

This paper utilizes an economic model to analyze the determinants of legislative decision making The model is empirically tested using the statistical technique of *n*-chotomous multivariate probit analysis The legislative issue addressed is the 1973 amendment to the Fair Labor Standards Act The amendment proposed to increase the minimum wage and the number of workers covered The dependent variable denotes a representative's voting pattern on the minimum wage, the independent variables represent economic characteristics of a legislator's congressional district The hypothesized linkages between the set of independent variables and legislative voting patterns were generally substantiated in the empirical tests

I. Introduction

The science of economics is concerned with the allocation of scarce resources among competing ends where the market system, central planning, and the political process are alternative institutional frameworks within which allocative decisions may be formulated Traditionally, economic theorists have concentrated on analyzing the workings of the market system and the operations of central planners Thus, the study of the political (collective decision making) process has been delegated to other disciplines Since the late 1940s, however, a body of literature has developed that attempts to integrate the collective decision-making

A preliminary version of this paper was presented at the annual meetings of the Atlantic Economic Society, September 28, 1974 Financial support was provided by the Old Dominion University Research Foundation The comments of George J Stigler, Gordon Tullock, John Knapp, Wayne Talley, Donald Snyder, and the referees of this journal are acknowledged with thanks

[*Journal of Political Economy*, 1976, vol 84, no 2]
© 1976 by The University of Chicago All rights reserved

may have raised common labor rates in low-wage areas faster than they would have risen otherwise, exerting some pressure also on rates above the minimum Thus, union influence in these areas has been exerted mainly through legislative pressure rather than through collective bargaining In hearings before the House Education and Labor General Subcommittee, March 1973, Andrew J Biemiller, director of the department of legislation of the AFL-CIO, urged favorable action on the minimum wage He challenged as "simply not true" arguments that the minimum wage caused higher unemployment for young people "The last few years have reduced the buying power of the minimum wage so that the minimum wage worker is worse off today than he was in 1966 when the minimum wage was $1 25 an hour " Biemiller recommended a uniform minimum wage for all workers of at least $2 immediately, with substantial increases in the next 2 years directed toward a $2 50 minimum "in the near future "

2 Regional impact (REG) The impact of the minimum wage is uneven It has virtually no direct effect on durable-goods manufacturing, construction, or public utilities It does affect labor-intensive branches of manufacturing, including textiles, canning, paper, and furniture The effect on these industries includes not only the wage impact but also industrial-location shifts Marshall Colberg, in his presidential address to the Southern Economic Association, hypothesized that " although studies of the effects of federal minimum wages have emphasized the consequent short-run losses in employment, it is quite certain that the long-run effects on the location of labor-intensive activities are more important These locational effects provide the primary political impetus for such legislation" (Colberg 1963, p 165) Colberg suggests that legislators from more industrialized areas support the minimum wage as a device to reduce the normal tendency of industrial shifts from their own relatively high-wage area to competitively lower-wage regions Additional support for the "Colberg hypothesis" is given by the following quotation from Senator Jacob Javits of New York "I point out to senators from industrial states like my own, that a minimum wage increase would give industry in our states some measure of protection, as we have too long suffered from the unfair competition based on substandard wages and other labor conditions in effect in certain areas of the country—primarily the south " On the basis of the industrial-location impact of the minimum wage, it is anticipated that labor-intensive manufacturing voter-consumers in high-wage (low-wage) areas would demand positive (negative) PSE on the minimum wage

3. Small business (SBI) The minimum wage adversely affects many small businesses, particularly the retail and service industry groups, by raising the cost of labor The 1973 legislation extended minimum-wage coverage gradually to approximately 700,000 retail and service employees

previously exempted. Thus, business leaders in these industry groups are expected to demand negative PSE on the minimum wage In hearings before the House Education and Labor General Subcommittee, March 1973, Robert T Thompson, in behalf of the U.S Chamber of Commerce, warned that the minimum-wage increases " . constitute a clear danger of creating greater unemployment and higher inflation." James R. Williams, president of the National Retail Merchants Association, warned that further increases in the minimum wage would mean curtailed hiring, lower employment levels, and a continuing shift toward employment of part-time workers A jump in the minimum wage would place an "intolerable burden" on the retail industry, Williams testified. Frederick L Williford, director of government affairs for the National Federation of Independent Business, testified that 84 percent of the members of his 335,000-member organization participating in a poll favored retaining the minimum wage at $1 60 In addition to the impact of the minimum wage on labor cost, small business and nonunion interests generally are expected to demand negative PSE on the minimum wage due to the adverse income effect That is, these voter-consumers are likely to be nearer average earnings which are well above any proposed legal minima Since the prices of goods and services they buy are increased following minimum-wage increases, their real income is reduced

4 Low-wage workers (LWW) Workers with wages below the proposed minimum may either benefit or lose If the minimum wage is established above the market rate, employment will decrease but wages for those still employed will rise As the minimum wage rises relative to the market rate, however, more and more labor units will find it in their interests to oppose such an increase in the minimum wage The higher the minimum wage relative to the market rate, the lower the ratio of those who benefit to those who are hurt and consequently the less the influence of those who benefit (Steindl 1974) The expected preference by this voter-consumer group will depend on the ratio of the income effect to the unemployment effect

5 Teenage workers (TEW). Teenage workers are expected to demand negative PSE on the minimum wage Several recent studies have concluded that the minimum-wage system has a particularly adverse effect on the employment of young people and has contributed to a rise in the teenage unemployment rate.[8] A legal minimum wage, furthermore, is detrimental to teenage workers because of their low productivity due to limited training and education The most controversial issue raised during the debate on the 1973 minimum wage involved the question of a youth-differential wage The youth-differential provision would have allowed employers to hire 16- and 17-year-old youths for 6 months at a

---

[8] For an example of one recent study, see Adie (1973).

subminimum wage In his veto message, President Nixon said the failure to provide a youth differential would cause increased unemployment among youths.[9]

The analysis in Section II can now be summarized by substituting the delineated beneficiaries and losers on the minimum wage into a political support effort function, that is, the level of political support effort on the minimum wage (PSE^*_{mw}) supplied by a legislator depends on the intensity of preferences by various voter-consumer groups

$$PSE^*_{mw} = F(\text{UNI, REG, LWW, SBI, TEW}) \qquad (1)$$

The political support effort function, PSE^*_{mw}, is continuous It is not limited to the exclusive alternatives of positive support versus zero support It is, instead, related to a level of effort by a representative—that is, the degree of lobbying and other supportive coercive activities engaged in The larger the PSE^*_{mw}, the greater the support effort by a representative for a legal minimum that is higher than the market rate and increased coverage Fairly weak PSE^*_{mw} could imply some support effort but for a legal wage that was not much different than the market rate and reduced coverage Even smaller PSE^*_{mw} might imply no positive support effort by a legislator or negative action on the minimum wage The independent variables specified above denote the delineated beneficiaries and losers under the minimum wage Based on the previous discussion, the sign of the UNI variable is predicted to be positive while the direction of influence for SBI and TEW is expected to be negative The sign for REG will depend on the industrial-location effects of the minimum wage in a congressional district, and the sign for LWW will be a function of the relative strength of the beneficiaries (income effect) and losers (unemployment effect) within this voter-consumer group

III. Empirical Method and Results

Legislative action to amend the Fair Labor Standards Act in 1973 is the observed basis for the empirical investigations The 1973 legislation passed by the House (HR 7935) increased the hourly minimum wage, extended coverage to about 6 million workers, and allowed overtime pay coverage to certain employees previously exempted. The final version of the legislation passed by the House was substantially the same as approved by the House Education and Labor Committee. The Erlenborn substitute, rejected by the House 199 to 218, differed greatly in scope and intent. The Erlenborn substitute would have stretched out the increase in the minimum wage to 3 years, continued the differential between farm and nonfarm workers, provided no new coverage, and permitted a youth differential Thus, the final House bill was much stronger than the Erlenborn substitute

---

[9] President Nixon's veto of HR 7935 was sustained by a 259 to 164 vote.

The roll-call voting pattern of an individual legislator on the Erlenborn substitute bill and the final House bill is used as a proxy variable to represent the optimum level of political support effort on the minimum wage Three categories of PSE^*_{mw} can be delineated·

Category one (YN) A legislator exhibiting this voting pattern is interpreted to be furnishing PSE on a restricted minimum-wage bill ("yes" on the substitute bill, "no" on the House bill) The YN category represents maximum effective observed opposition on the 1973 minimum-wage amendments[10]

Category two (YY) A legislator adopting this voting pattern appears to be willing to provide PSE on any minimum-wage bill ("yes" on the substitute, "yes" on the House bill)

Category three (NY) A legislator following this voting pattern is supplying PSE for a comprehensive minimum-wage bill ("no" on the substitute bill, "yes" on the House bill)

The three observed categories of legislative roll-call voting behavior represent levels of PSE on the minimum wage which increase from category one through category three These categories denote three observations of an underlying continuous dependent variable, PSE^*_{mw}, which cannot be empirically measured[11]

Since the observations on the dependent variable are categorical rather than continuous, multiple regression analysis is not an appropriate empirical technique[12] Therefore, *n*-chotomous multivariate probit analysis is used in the investigations which follow[13] The number of observations used was 403 Observations omitted were from vacant congressional districts and districts with a representative not voting on the

---

[10] The NN class was not empty Five representatives followed a NN voting pattern, these were Idaho (2), Indiana (2), North Dakota, Texas (15), and Virginia (8) We felt that this small sample (1 percent of the representatives) would not lead to meaningful empirical results The NN pattern, moreover, was not considered relevant because there existed a consensus in Congress that a minimum-wage increase was necessary, the debate concerned the extent of coverage and the time pattern of the increase

[11] The organization of the independent variable must follow the Guttman scaling method Assigning a zero (0) for a negative response and one (1) for a positive response, the three categories follow a perfect Guttman scale 00, 10, 11 An explanation of the Guttman technique may be found in Mokken (1971)

[12] Serious statistical problems may arise when multiple regression is used to analyze data with ordinal observations on the dependent variable First, it cannot be assumed that the expected value of the error term is zero Second, the assumption that the error term has a constant variance cannot be met Third, the error term is not distributed normally. Fourth, regression analysis applied to ordinal data makes the unrealistic assumption that the data are cardinal

[13] Multivariate statistical models that deal with a dichotomous dependent variable are discussed in Tobin (1958) and Goldberger (1964) The dichotomous multivariate probit models have been extended to handle more than two categories of the dependent variable by Zavonia and McKelvey (1969) and Johnson (1972)

TABLE 1

1973 BREAKDOWN OF THE DEPENDENT VARIABLE

Code Value of Response	Frequency of Observation
R(YN) = 1 0	119
R(YY) = 2 0	75
R(NY) = 3 0	205

minimum wage[14] A breakdown of the dependent variable is reported in table 1

Due to measurement difficulties, the proxy variables used to measure the intensity of preferences of the voter-consumer groups specified in equation (1) are the following (1) UNI—the total campaign contributions of organized labor received bv the winning candidate in the 1972 congressional election is used to measure the degree of political participation by labor union interests,[15] (2) REG—a dummy variable was used taking on the value 1 if a congressional district is located in the South and 0 if the district is located elsewhere This variable, once account is taken of the other independent variables, should isolate the negative preferences on the minimum wage in the South due to concern over losing their low-wage competitive advantage in labor-intensive industries over other regions,[16] (3) SBI—the total campaign contributions from small business organizations received by the winning candidate in the 1972 congressional elections is used to measure the degree of political participation by these voter-consumers,[17] (4) LWW—the percentage of families and unrelated individuals over 14 years of age earning less than $3,999 is used as a proxy for LWW The larger the percentage of workers earning less than the proposed minimum, the greater the number of voter-consumers affected by the legislation, and (5) TEW—the percentage of the total

---

[14] An abstain vote or the absence of a legislator was not included as a possible category of the dependent variable because only a few observations followed this pattern With only a few abstain votes and absences, the reasons for the action may not conform to meaningful legislative action Consider, for example, the possibility of illness or other personal problems prohibiting a legislator from being present at the time of the voting

[15] The data source for campaign contributions by both labor and business was Common Cause (1974).

[16] We expect that the region dummy variable would not matter as much if a good measure of wage imports were available A measure of the relative wage in labor-intensive industries would represent an ideal proxy variable Data to construct such a wage measure are not available by congressional districts We felt that the use of a South/North breakdown was justified in order to capture the "Colberg effect " Furthermore, only southern congressional districts were separated out due to the documented industrial-location shifts of labor-intensive industries to the South because of the low-wage, nonunion advantages

[17] Small business organizations included several businessmen's groups, primarily the National Association of Manufacturers

TABLE 2

N-CHOTOMOUS MULTIVARIATE PROBIT ANALYSIS: THE OPTIMUM LEVEL OF POLITICAL SUPPORT EFFORT ON THE 1973 MINIMUM-WAGE LEGISLATION

Variable	Maximum-Likelihood Estimate*	β Coefficient
Labor unions (UNI)	0 00029†	1 47235
Small business (SBI)	−0 00016†	−0 26990
Regional impact (REG)	−0 61962†	−0 28917
Low-wage workers (LWW)	1 75127‡	0 17466
Teenage workers (TEW)	−12 09548†	−0 19570

* Minus 2 times the log of the likelihood ratio = 227 31528, statistically significant at the 1-percent level or better, χ^2 distribution with 5 df

† Statistically significant at the 1-percent level or better, χ^2 distribution with 1 df

‡ Statistically significant at the 5-percent level, χ^2 distribution with 1 df

population that is 16–21 years of age is used as a proxy to account for teenage workers

The results of the empirical testing are reported in table 2 The equation is statistically significant at the 1-percent level or better, as measured by the likelihood-ratio test[18] The generalized likelihood-ratio test was used because the exact properties of the variance-covariance matrix are not known Thus, the sampling-distribution test of significance (*t*-test) may not be accurate For the likelihood-ratio test, the estimation is first performed with the β's unrestricted as reported in table 2 It is known that −2 times the log of the likelihood ratio (λ^*) represents the test statistic for overall significance To test the significance of individual coefficients, the estimation is performed with the variable under investigation omitted, ($\lambda^{*\prime}$) denotes the corresponding test statistic with the variable under question excluded The fact that ($\lambda^* - \lambda^{*\prime}$) is distributed as χ^2 with 1 df was used as the test statistic for each variable. The critical region for the likelihood-ratio test is the right-hand tail of the χ^2 distribution.[19]

The test for significance revealed that the UNI, SBI, REG, and TEW variables are statistically significant at the 1-percent level or better. These variables have the predicted direction of influence on explaining the variation in the dependent variable Each variable, moreover, added significantly to the combined explanatory power of the equation. The LWW variable, for which the sign was not specified, exerts a positive influence This indicates that workers earning below the proposed

---

[18] A statistical test that measures the explained variation in the dependent variable is not reported The dependent variable for the empirical tests is categorical and represents the "true" dependent variable, which cannot be empirically measured A discussion of the differences between multiple regression and *n*-chotomous probit with regard to interpretation may be found in Silberman and Talley (1974)

[19] A thorough discussion of the generalized likelihood-ratio test may be found in Mood and Graybill (1963)

TABLE 3

CORRELATION COEFFICIENTS FOR VARIABLES IN 1973 MINIMUM-WAGE EQUATION

	UNI	SBI	REG	LWW	TEW
UNI	1.000	−0.213	−0.178	−0.117	−0.118
SBI		1.000	0.019	−0.118	−0.007
REG			1.000	0.304	0.149
LWW				1.000	0.531
TEW					1.000

minimum wage demand positive PSE_{mw} because the relative intensity of preferences of the expected beneficiaries under the proposed legislation outweigh the expected losers.[20] That is, LWW support the minimum wage due to the perceived beneficial income effect.

An examination of the correlation matrix $(r_{x_i x_j})$ as well as the overall significance of the equation reveals that multicollinearity does not appear to be a serious statistical problem. The correlation coefficients are presented in table 3. The LWW variable, however, is correlated with TEW to a relatively high degree (.531). The influence of LWW was investigated by deleting it from the equation and then inserting it in a subsequent equation. The LWW variable improved the overall statistical significance without rendering the individual coefficients unacceptable ("wrong") on a priori considerations. Thus, the LWW variable was considered useful and was retained as an explanatory variable.

In an investigation into the optimum level of political support effort on the minimum wage, it would be informative to compare and contrast the relative importance of the independent variables. Since the maximum-likelihood estimates are in different units, they cannot be compared directly. To overcome this, the coefficients may be converted into units of standard deviations.[21] The results of these calculations are reported in table 2.

The most important variable was UNI. It is not surprising that union interests had the largest influence. Unions felt intensely about passage of a strong minimum-wage bill and contributed significant amounts of campaign funds to representatives. The REG variable was next in relative importance, followed by SBI. Small business interests, although having strong negative preferences on the minimum wage, are not as politically organized as labor interests. The importance of the region dummy variable lends support to the "Colberg hypothesis" delineated previously. The LWW and TEW variables were not relatively important This may be attributed to two factors the fact that their stake in the minimum wage is not as important as other voter-consumer groups and political incompetence That is, these voter-consumer groups do not possess intense preferences on the minimum wage, and fewer of them vote and contribute campaign funds

---

[20] Our results do not invalidate the hypothesis that the lower the market wage relative to the proposed minimum the larger the probability of being injured by the law. The positive LWW coefficient suggests that in most congressional districts, the intensity of preferences of voter-consumers who feel threatened is not enough to get their representative to furnish PSE against the minimum wage. A legislator will not change his voting pattern from "for" to "against" until the intensity of preferences for those perceived to be helped is less than the intensity of preferences for those perceived to be hurt.

[21] The methodology for converting partial coefficients into units of standard deviations may be found in Yamane (1969).

IV. Conclusions

In recent years, considerable attention has been given to the development of a theory of collective decision making However, relatively little work of an empirical nature has been generated This condition appears to have at least three primary causes (1) existing models have failed to link theory with quantifiable variables, (2) data have not been readily available, and (3) appropriate analytical techniques have not been properly utilized In this paper, we have taken what we consider to be a small but important remedial step which may lead the way to a deeper understanding of the collective decision-making process Specifically, a model of legislative decision making was developed and tested using the technique of *n*-chotomous multivariate probit analysis

The results obtained were consistent with the theory The independent variables were statistically significant with proper signs in all cases Unions, small business, and region were the most important, while low-wage and teenage workers also appeared to influence roll-call voting patterns as expected Moreover, the results with respect to union and small business interests furnish significant statistical evidence regarding the impact of campaign contributions on the actions of political suppliers

In general, the study appears to have provided sufficiently useful and accurate results, based on theoretical expectations, to merit much more intensive and extensive work by ourselves and others However, certain problems remain to be solved in order to refine the techniques and methodology used here In some instances it may be difficult to define satisfactory proxies for specified intensities of preference by voter-consumer groups [22]

---

[22] For the independent variables in this study, the a priori relationships were based on traditional economic association reinforced by testimony gathered from the records of congressional hearings. However, a given model specification may be broader than existing data available for empirical tests The collection of primary data through survey would then be necessary before a complete set of parameter estimates could be generated. This shortcoming does not, of course, invalidate multivariate statistical analysis with a truncated variable set under the usual assumptions of independence. Thus, the use of relatively inexpensive data from census counts is a legitimate, if incomplete, exercise.

In summary, we believe that the study represents a positive and needed step toward empirical verification of existing economic models of collective decision making This is particularly so because the research begun here can readily be extended since minimum-wage legislation is periodically rewritten. In addition, there are a host of other legislative issues which can be similarly investigated, such as open housing, educational subsidies, vocational training, etc In this way, it can be determined whether our approach can generate knowledge of the legislative process which is general in nature The effects of campaign financing, moreover, can be carefully analyzed by relating contributions to the roll-call voting pattern of political-good suppliers

References

Adie, Douglas K "Teen-Age Unemployment and Real Federal Minimum Wage" *J P E* 81, no 2 (March/April 1973) 420–29

Black, Duncan *The Theory of Committees and Elections* Cambridge Cambridge Univ Press, 1958

Breton, Albert *The Economic Theory of Representative Government* Chicago Aldine, 1974

Buchanan, J., and Tullock, G *The Calculus of Consent* Ann Arbor Univ Michigan Press, 1962

Cherryholmes, C. H , and Shapiro, M J *Representatives and Roll Calls.* New York Bobbs-Merrill, 1969

Clausen, R. A *How Congressmen Decide A Policy Focus* New York St Martin's, 1973

Colberg, Marshall R "Human Capital as a Southern Resource" *Southern Econ J* 29, no 3 (January 1963) 157–65

Coleman, James S *The Mathematics of Collective Action* Chicago Aldine, 1973

Common Cause "1972 Congressional Finances" Xerox, 1974

Downs, Anthony *An Economic Theory of Democracy* New York Harper & Row, 1958.

Dye, Thomas *Politics, Economics and the Public* Chicago Rand-McNally, 1966

Frohlich, N , Oppenheimer, J , and Young, O *Political Leadership and Collective Goods.* Princeton, N J Princeton Univ Press, 1971

Goldberger, Arthur *Econometric Theory* New York Wiley, 1964

Johnson, Thomas "Qualitative and Limited Dependent Variables in Economic Relationships" *Econometrica* 40 (May 1972) 455–62

Mokken, R J. *A Theory and Procedure of Scale Analysis* The Hague Maton, 1971.

Mood, Alexander M , and Graybill, Franklin A *Introduction to the Theory of Statistics.* New York: McGraw-Hill, 1963

Olsen, Mancur, Jr *The Logic of Collective Action.* Cambridge, Mass.. Harvard Univ. Press, 1965

Rothenberg, Jerome "A Model of Economic and Political Decision Making." In *The Public Economy of Urban Communities*, edited by Julius Margolis. Baltimore· Johns Hopkins Press, 1965

Silberman, Jonathan, and Talley, Wayne K., "*N*-chotomous Dependent Variables: An Application to Regulatory Decision-Making." In *1974 Business and Economic Statistics Section, Proceedings of the American Statistical Association.* Washington. American Statistical Association, 1974.

Steindl, Frank G. "The Appeal of Minimum Wage Laws and the Invisible Hand in Government." *Public Choice* 14 (Winter 1974): 133–36.

Tobin, J "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26 (1958) 24–36.

Yamane, Taro *Statistics An Introductory Analysis*. New York. Harper & Row, 1969.

Zavonia, William, and McKelvey, Richard "A Statistical Model for the Analysis of Legislative Voting Behavior " Paper presented at the annual meetings of the American Political Science Association, New York, 1969

Communications

Perceived Wealth in Bonds and Social Security: A Comment

Martin Feldstein

Harvard University

In a recent paper, Robert Barro (1974) presented a model in which government debt and social security are not perceived by households to add to net wealth From this it follows that neither government debt nor social security affects private capital accumulation These important results contradict the modern theory of government debt[1] and recent contributions to the analysis of social security.[2]

Barro's emphasis on bequests is an important extension of the traditional life-cycle model Nevertheless, the existence of such voluntary intergenerational transfers does not have the striking implication that Barro asserts I will show in this note that Barro's conclusions follow instead from restrictive and empirically unwarranted assumptions In a more appropriate model that includes bequests, the introduction or increase of government debt or of social security will reduce private savings and the equilibrium ratio of capital to labor

The framework of Barro's analysis is a static intertemporal model with overlapping generations of the type first presented by Samuelson (1958),

---

Support by the National Science Foundation is gratefully acknowledged

[1] Buchanan (1958), Modigliani (1961), and Diamond (1965) present models in which public debt is regarded as net wealth by the households so that real capital accumulation is depressed Unlike Barro, these authors disregard the role of bequests. An analysis similar to Barro's is presented in Kochin (1974) and Miller and Upton (1974)

[2] I discussed the theory and evidence on the relation between social security and private savings in Feldstein (1974*a*, 1976) with a model that ignores bequests. If social security does not alter retirement behavior, it will reduce private savings. If additional retirement is induced, the effect of social security on savings is a priori uncertain. The empirical evidence strongly suggests that social security does reduce savings

[*Journal of Political Economy*, 1976, vol 84, no 2]
© 1976 by The University of Chicago All rights reserved

A novel feature of Barro's model is that each individual's utility depends not only on his own consumption but also on the utility of his immediate heir The utility of the current generation therefore depends indirectly on the utility of all future generations The introduction of government debt creates wealth for the current generation but a liability for a future generation or generations The current generation will resist this redistribution from their hiers to themselves by increasing their bequests by just enough so that their heirs are no worse off The extra saving for this additional bequest just offsets the reduced savings that would otherwise be associated with the government debt Barro summarizes this ingenious argument by writing, "However, if, prior to the government bond issue, a member of the old generation had already selected a positive bequest, it is clear that this individual already had the option of shifting resources from his descendant to himself, but he had determined that such shifting, at the margin, was nonoptimal Since the change in [government debt] does not alter the relevant opportunity set in this sense, it follows that—through the appropriate adjustment of the bequest—the values of current and future consumption and attained utility will be unaffected" (p 1103)

There is an analogous effect of social security in this model Social security is a forced transfer from the next generation to the current generation If the current generation is in equilibrium with positive bequests when social security is introduced, individuals will increase their bequests by just enough to pay their descendants for the social security taxes they are forced to pay In Barro's words, "As in the case of changes in government debt, if the solutions for bequest are interior, the impact of a marginal change in [social security] would be solely on the size of bequests and not at all on the pattern of consumption" (p 1107)

These conclusions reflect the crucial assumption in Barro's analysis that the economy has a constant population and no economic growth Consider instead an economy in which total national income grows at rate g, the sum of the growth rate of population and the rate of technical progress It is useful to analyze first the special case in which the rate of interest on government debt (r) is not greater than the rate of economic growth [3] In this circumstance, the government can create debt and yet never have to levy a future tax to repay the debt or to pay interest on the debt Instead, the government merely issues new debt with which to pay the interest The debt therefore grows at the rate of interest r Since $r \leq g$, the ratio of debt to national income will not increase. The first generation —that is, the generation that receives the debt as a transfer from the government—knows that no future generation will be called upon to pay the debt There is no need, therefore, to increase the previously planned bequests The first generation will therefore increase its own consumption and thus reduce capital accumulation

---

[3] Samuelson (1958) considered the impact of social security in the extreme case in which $r = 0$ but $g > 0$ Cass and Yaari (1967) discuss social security in the current case of $g > r > 0$ Neither of these studies allowed for bequests Barro suggests that $r < g$ is inconsistent with a model of rational intergenerational bequests This does not preclude $r = g$ Moreover, $r < g$ can be maintained if each generation forgoes the opportunity to reduce the capital stock in order to allow the next generation to have the opportunity

The result is the same for social security In steady-state growth, the benefits that an individual receives in a pay-as-you-go social security system exceed his tax contributions More specifically, the individual's tax contributions earn an implicit "rate of return" that is equal to the rate of growth of the tax base, that is, the rate of growth of national income Since $g \geq r$, each individual receives a "rate of return" on his social security tax payments that is not lower than he would have received by investing that same amount Thus, although the heirs of the first generation are compelled to pay taxes to support the first generation members in their old age, these taxes do not represent a net loss to the second generation since they will lead to subsequent benefits with a present value that is at least as great There is no need, therefore, for the first generation to increase previously planned bequests Indeed, if subsequent generations are all better off, the first generation can reduce its planned bequest In addition, the present value of lifetime consumption will increase by the present value of the social security benefits, and capital accumulation will be reduced accordingly

The assumption that $g \geq r$ is analytically convenient but empirically false[4] Consider, therefore, the case of $r > g$ Although r is the rate of interest that the government pays on its debt, r may overstate the net cost of the debt If the government taxes interest payments at rate θ, the net interest cost per dollar of debt is $r_N = (1 - \theta)r$ For realistic values of θ and r, the economy may be characterized by $r > g \geq r_N$ The property that $g \geq r_N$ makes this case very similar in its behavioral implications to the previous situation of $g \geq r$ Again, the interest on government debt can be financed by issuing new debt with the total debt growing at the rate $r_N \leq g$ Similarly, social security provides a higher implicit rate of return (g) than the rate r_N that descendants could otherwise earn,[5] so that they are better off even without an additional bequest than they would have been in the absence of social security

---

[4] This statement is correct if r is interpreted as the marginal product of capital The actual long-term rate of interest on the government debt in less inflationary times was less than 0 04 and therefore less than the rate of growth of real national income The government interest rate is still less than the relevant rate of growth of money national income

[5] In a perfect capital market, all assets have a yield that is equivalent, after adjustment for risk, to r In practice, the vast majority of households invest almost exclusively in savings accounts which yield less than r

Even if $r_N > g$, the creation of public debt or social security "wealth" will still depress savings in a growing economy. If $r_N > g$, some future taxes would have to be levied to limit the growth of the government debt. Similarly, the social security program would actually lower the welfare of future generations Nevertheless, both the required tax and the inefficiency of social security would be small if $r_N - g$ is small The first generation will therefore increase its bequests by less than the value of the debt or social security transfers that it receives For example, if accumulated interest makes the debt grow at the rate of growth of the economy, taxes will be required to finance only the fraction $1 - g/r_N$ of future interest payments The first generation will therefore increase its bequest only by the fraction $1 - g/r_N$ of the debt that it is given

Barro touched briefly on the possibility of a growing economy in a footnote on page 1105 He noted first that his own line of proof depends on the assumption that the government debt will eventually be repaid and that the social security scheme will eventually be liquidated He argues correctly that this assumption is not crucial in a static economy. He then tries to extend this to a growing economy His argument about the government debt[6] rests on the fact that the debt cannot grow faster than national income He then incorrectly asserts that this limit precludes financing all interest payments by issuing new debt whenever $r > g$ From this he concludes that the ability to issue new debt in a growing economy is irrelevant to his conclusion As I noted above, net interest costs can be financed by creating new debt whenever $r_N < g$ even if $r > g$ Moreover, even if $r_N > g$, the fraction g/r_N of each year's interest payments can be financed by additional debt creation The fraction g/r_N of the debt need never be financed by taxation and thus represents net-wealth creation to the infinite sequence of households Only in the special case of a static economy ($g = 0$) is the present value of required future taxes equal to the present value of the debt itself

The careful reader may well question whether it is appropriate to regard the net cost to the government of interest payments to be $(1 - \theta)rD$, where D is the size of the government debt If an increase in the government debt reduces private capital accumulation, the government will forgo the taxes on the income produced by that capital If the reduction in private capital is equal to the addition to public debt, and if private capital earns a marginal rate of return r, the lost tax income is θrD so that the total annual cost to the government of debt D is $(1 - \theta)rD + \theta rD = rD$[7] In this case, the effective net rate of interest is r, not $(1 - \theta)r$.[8] The conclusion that the effective rate of interest is r does not, however, eliminate the impact of social security or public debt. The government can still finance the fraction g/r of its interest payments by additional debt creation, making this same fraction of the original debt an increment to wealth

---

[6] No specific argument is given about social security. Barro suggests (p. 1107) that the previous argument about repaying the debt is relevant to social security.

[7] In practice, the issue is far more complex because of the existence of the corporation tax, equity finance, and the deductibility of interest payments by corporations. The above analysis is appropriate only if marginal corporate investments are financed by debt creation

A similar argument applies to social security Although individuals may find that social security provides a greater implicit rate of return (g) than the net rate of interest that they can earn $[(1 - \theta)r]$, the government must raise additional taxes to replace the tax revenue that is lost when social security depresses private savings Future generations are therefore worse off because of social security if $r > g$ But the present value of this reduction in future welfare is not as great as it would be in a static economy or if the program would eventually be liquidated The original creation of social security should not, therefore, induce the first generation to increase its bequests by the amount of the benefits that it receives

There is a further complication if the government debt or household "social security wealth" is large in relation to the capital stock The induced reduction in capital accumulation will then lower the wages in future generations The first generation should anticipate this by increasing its bequests to compensate for the future fall in earned income

The complexity of these anticipations casts doubt on the empirical relevance of this entire exercise In determining their bequests, households are required to understand the effect of these bequests and of the financing requirements of the public debt and the social security program on all future generations They must recognize that any reduction in saving will reduce tax revenues and therefore require other tax increases They must understand the general equilibrium effect of reduced saving on future wages The subtlety of the required anticipations makes it unlikely that households actually do respond as Barro's model suggests Since the economics profession has previously ignored the need to adjust bequests to compensate for future tax liabilities, it may be safe to assume that households have also not made the adjustment implied by the new theory

It is also clear that for most families the voluntary and intentional "intergeneration transfers" are not bequests at death but the support of the consumption that their heirs enjoy as children. Some of this consumption may be the purchase of education or other investment in human capital The nature of these transfers makes it even less likely that households will respond as the theory suggests Moreover, even if households do respond to an increase in government debt or social security by increasing such intergeneration transfers, this merely changes the nature of the induced consumption and does not constitute a transfer of real capital.

---

[8] There may be no loss in tax revenue if the economy is open to foreign investment, and international capital flows equate the rate of return in all countries In this case, any reduction in domestic saving will be offset by a capital inflow from abroad Whether this replaces the lost tax revenue on domestic interest income depends on the tax rules and the mix of debt and equity finance.

In short, the introduction of social security or public debt in a growing economy can be expected to reduce the accumulation of real capital. Moreover, since the accumulation of a trust fund by the social security program is equivalent to a reduction in public debt, such a fund would increase capital accumulation [9]

References

Barro, Robert J "Are Government Bonds Net Wealth?" *J P.E* 82, no. 6 (November/December 1974) 1095–1117

Buchanan, James. *Public Principles of Public Debt* Homewood, Ill. Irwin, 1958.

Cass, David, and Yaari, M F "Individual Saving, Aggregate Capital Accumulation, and Efficient Growth " In *Essays in the Theory of Optimal Growth,* edited by K Shell Cambridge, Mass M I T. Press, 1967

Diamond, Peter A "National Debt in a Neoclassical Growth Model " *A E.R.* 60 (December 1965) 1126–50.

Feldstein, Martin S "Social Security, Induced Retirement, and Aggregate Capital Accumulation " *J P E* 82, no 5 (September/October 1974) · 905–26. (*a*)

——— "The Optimal Financing of Social Security " Harvard Institute of Economic Research Discussion Paper no 388, November 1974 (*b*)

——— "Social Security and Private Savings International Evidence in an Extended Life Cycle Model." In *The Economics of Public Services,* edited by M Feldstein and R Inman London Macmillan, 1976, in press

Kochin, L "Are Future Taxes Anticipated by Consumers?" *J Money, Credit and Banking* 6 (August 1974). 385–94.

Miller, M H, and Upton, C W *Macroeconomics A Neoclassical Introduction* Homewood, Ill Irwin, 1974

Modigliani, Franco "Long-Run Implications of Alternative Fiscal Policies and the Burden of the National Debt " *Econ J.* 71 (December 1961) 730–55.

Samuelson, Paul A "An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money." *J P.E* 66, no 6 (December 1958) 467–82

---

[9] Feldstein (1974*b*) discusses the desirability of such a fund

Barro on the Ricardian Equivalence Theorem

James M. Buchanan

Virginia Polytechnic Institute and State University

Is public debt issue equivalent to taxation? This is an age-old question in public finance theory David Ricardo presented the case for the affirmative [1] Professor Robert J Barro reexamines the question in his recent paper (1974) without, however, making reference to Ricardo or other early contributors Although his discussion is carefully qualified to allow for exceptions under specified conditions, the thrust of Barro's argument supports the Ricardian theorem to the effect that taxation and public debt issue exert basically equivalent effects.

Barro's central emphasis is on demonstrating that, under reasonable conditions which involve overlapping generations of persons with finite lives, taxpayers will capitalize the future obligations that public debt issue embodies To the extent that this capitalization occurs, government bonds do not add to the perceived net wealth in the economy. From this Barro infers that the substitution of debt for tax finance will exert no expansionary effect on total spending There are two questions here Are the future tax liabilities fully capitalized? And, even if they are, does this necessarily imply that the fiscal policy shift exerts no effect on total spending? To establish the second result, it is necessary to examine the differential impacts of taxation and debt issue, quite apart from the question of the capitalization of future taxes Barro wholly neglects this necessary part of any comparative analysis of the two fiscal instruments, and, because of this neglect, his conclusion is not nearly so relevant for policy as it seems to be This neglect may stem from Barro's failure to specify properly the inclusive set of transactions that debt issue represents

---

I am indebted to my colleagues Gordon Tullock, Nicolaus Tideman, and Richard Wagner for helpful comments

[1] The basic Ricardian discussion is contained in Ricardo (1951, vol 1, pp 244–49) and "Funding System" (vol 4, pp 149–200) Antonio De Viti De Marco elaborated the Ricardian thesis in a somewhat modified setting His analysis was first developed in De Viti De Marco (1893) Essentially the same analysis appears in De Viti De Marco (1936, bk 5, chap 1, pp 377–98) For my own summary discussions, see Buchanan (1958, pp 43–44, 114–22).

[*Journal of Political Economy*, 1976, vol 84, no 2]
© 1976 by The University of Chicago. All rights reserved

After constructing his overlapping-generations model with no government sector, Barro then superimposes an issue of public debt without offsetting or compensating changes. He states, "Suppose now that the government issues an amount of debt, B, which can be thought of as taking the form of one-period, real-valued bonds It can be assumed, for simplicity, that the government bond issue takes the form of a helicopter drop to currently old (generation 1) households" (p 1101) We are immediately prompted to inquire why any government would undertake such an activity If the purpose is to increase total spending in the economy, this might be done much more readily by the straightforward issue of money, which does not involve future payment obligations Or, if the purpose is to finance public spending without current taxation or money creation, bonds must be sold in the private capital market, and persons who purchase these bonds must draw down private investments or reduce private consumption Under this condition, it is impossible that "the increase in B implies a one-to-one increase in the asset supply " (p 1103)

There are two alternative interpretations of Barro's seemingly bizarre model of bond issue The first and more constructive interpretation suggests that Barro introduces the model only because of his concentrated interest on the wealth effect of debt issue within the strictly Ricardian or differential-incidence framework The second interpretation suggests that Barro's analysis applies only when governments do, in fact, incur future payment obligations without securing command over funds in the initial period, a setting which does not apply to public debt as ordinarily conceived but which may apply to the operation of the social security system in the United States

These quite separate interpretations may be discussed briefly in turn Barro directly follows the first statement cited above with the following "Equivalently, it could be assumed that the bonds were sold on a competitive capital market, with the proceeds of this sale used to effect a lump-sum transfer to generation 1 households" (p 1101) Superficially, the two operations are not at all equivalent If, however, the bonds dropped from the helicopter are marketable, those who get them may, as desired, convert them into currency The net effects are the same as the government's sale of bonds to voluntary purchasers along with the lump-sum transfer of the proceeds to the same persons who would have received the bonds in the first model In either case, the government bonds will replace private bonds or other assets, and the net-wealth effects of this substitution must be considered along with the wealth effects embodied in the capitalization of future tax liabilities Why might Barro have felt that he could neglect the first of these considerations? He could have done so if his interest is exclusively on determining the differential effects of

(a)

Taxation

Assets	Liabilities
Public assets (+)	
Private assets (-)	

(b)

Debt Issue

Assets	Liabilities
Public assets (+)	
Private assets (-)	
- - - - - - - - - -	- - - - - - - - - -
Government bonds (+)	Present value of tax liabilities (+)

FIG 1

debt and tax financing, and he is willing to make the severely restrictive assumption that the source for the ultimate purchase of the government bonds is identical to the source from which the alternative taxes would be drawn

This may be clarified by the following simple *t*-account comparison Under the tax alternative, the *t*-account adjustments appear as in figure 1*a* Under either of the debt-issue alternatives, the final *t*-account adjustments are as shown in figure 1*b* If we are interested only in comparing the differential effects of tax and debt finance for a given level of spending, and if we are willing to make the heroic assumption that the direct impact of these two fiscal instruments is identical, then the part of figure 1*b* below the dotted line represents the net difference between the two combined fiscal operations

In this interpretation, Barro is simply not concerned either with the independent effects of debt-financed spending on net wealth or with the effects of public debt issue in a modified differential-incidence framework Does the financing of public spending by genuine public debt increase aggregate demand in the economy? This is a meaningful and relevant question that has been more widely discussed than the one which Barro analyzes In terms of figure 1, this question requires us to look at the adjustments in figure 1*b* only, without regard to any comparison with 1*a*. Provided that we do not go beyond the simple *t*-accounts, the net-wealth effect of debt financing of public outlay seems to depend on the degree to which future taxes are discounted But this neglects the possible effects of the sale of bonds on displaced private borrowers To the extent that these borrowers, those who would have secured the funds had not the government done so, would have applied the same discount factor to the future payment obligations as that applied by the public borrowers (the taxpayers), net wealth will not be modified regardless of the degree of capitalization of future taxes It seems reasonable to suggest that perceived net wealth, as relevant for spending behavior, increases with increases in either private or public debt But there are offsetting changes, and the direction of net effect depends on the differences [2]

Perhaps the most familiar setting is one that compares the effects of debt financing with explicit money creation for a given increment in public outlay This, too, is a differential-incidence model but one that is quite separate from the Ricardo-Barro model The analysis here requires all of the considerations noted above with respect to independent debt issue but has the advantage of emphasizing the specific deflationary impact of debt sales quite apart from the question of future tax capitalization [3] This latter aspect assumes primary importance, of course, when the principal objective for fiscal policy is to generate an expansion in aggregate spending

Barro's analysis may be interpreted in a second and more restrictive way that takes his explicit model for debt issue at face value and makes no attempt to generalize it toward meeting the expressed purpose of the paper In this view, Barro analyzes the issue of debt by a government which does not secure funds in the initial period and which does not allow its debt obligations to be resold in markets Something analytically analogous to this sort of operation may take place when legislation is enacted which commits the government to pay old-age pensions without at the same time fully financing such obligations with current levies of taxes Barro does, in fact, refer to the applicability of his analysis to social security financing, and he may have been led to generalize too readily from this somewhat singular example In this setting, the aggregate economic effects depend strictly on the relative discounting of future benefits on the one hand and future taxes on the other

---

[2] In a strict accounting sense, future obligations that are implicit in any debt issue, public or private, may be fully capitalized, while at the same time behavior may reflect the rational intent of borrowers to accelerate spending In this context, a lending-borrowing transaction must affect aggregate spending quite apart from the absence of change in measured net wealth See R N McKean (1951, particularly p 75)

[3] See Musgrave (1959, p 538) For particular emphasis on the deflationary impact of debt issue see Rolph (1957)

To this point my criticism of Barro's paper has been confined to the ambiguities in his setting for public debt issue along with the possible incompleteness of his analysis I have not challenged his basic conclusion that future taxes will tend to be fully discounted save under the exceptions that he notes Nonetheless, this conclusion may be questioned on empirical grounds What evidence might be adduced to suggest that individuals do not, in fact, modify their behavior fully to adjust for the present values of the future tax obligations that public debt issue implies?

Consider, first, the social security system which, as indicated, fits more closely with the narrow interpretation of Barro's results One implication of his analysis would be that the social security system as it has operated should not have modified the rate of private saving in the economy This runs counter to the result that has been observed by Martin Feldstein, who estimates that the rate of private saving has been reduced by some 38 percent [4] Additional evidence may be indirectly available from the historical record if the basic public choice paradigm is accepted If politicians are ultimately responsive to the desires of their constituents, we may infer something about constituents' evaluations by observing the behavior of politicians The 40-year history of social security financing yields ample evidence that politicians are extremely reluctant to adopt anything which smacks of full funding for the system Under the Barro hypothesis, there should be roughly indifferent public reactions to a fully funded and to an unfunded pension system

If we shift to the more general model, governments should be roughly indifferent as between financing current outlays from taxation and from genuine debt issue There should be no effect of debt-financed deficits on aggregate spending Direct testing of this hypothesis is difficult because of the fact that budget deficits are normally financed by some combination of genuine debt issue and money creation [5] Again, however, the behavior of legislators seems to offer indirect evidence against the capitalization hypothesis Can anyone in the post-Keynesian world of 1975 seriously question the proclivity of politicians to expand public debt in preference to tax increases? At the state-local level, where the money-creation powers do not exist, why should we observe constitutional constraints on public debts?[6] With public as with private finance, the very creation of debt suggests that borrowers desire to accelerate spending

---

[4] See Feldstein (1974) It should be noted that, even with full lifetime discounting of future taxes, a life-cycle model of utility maximization would produce a negative effect on private saving under a pay-as-we-go public pension system But, of course, the life-cycle model itself implies that debt financing necessarily differs from tax financing to the extent that debt is not retired within a generation

[5] Kochin (1974) attempted to estimate empirically the impact of deficit spending on private consumption His conclusions tend to support the capitalization view.

[6] Admittedly, the out-migration prospect for citizens in local jurisdictions offers an offsetting influence to full discounting

References

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82, no. 6 (November/December 1974) 1095–1117

Buchanan, James M *Public Principles of Public Debt.* Homewood, Ill.: Irwin, 1958.

De Viti De Marco, Antonio. "La Pressione tributaria dell' imposta e del prestito." *Giornale degli economisti* 1 (1893). 216–31

——— *First Principles of Public Finance* Translated by E. P. Marget. New York Harcourt Brace, 1936

Feldstein, Martin. "Social Security, Induced Retirement, and Aggregate Capital Accumulation " *J.P.E.* 82, no. 5 (September/October 1974) 905–26

Kochin, Lewis "Are Future Taxes Anticipated by Consumers?" *J. Money, Credit and Banking* 6 (August 1974) 385–94

McKean, R N. "Liquidity and a National Balance Sheet " *J P E* 57, no 6 (December 1949) 506–22 Reprinted in *Readings in Monetary Theory.* Edited by F A Lutz and L W Mints New York Blakiston, 1951.

Musgrave, R. A *The Theory of Public Finance.* New York McGraw-Hill, 1959.

Ricardo, David. *Principles of Political Economy and Taxation, Works and Correspondence* Edited by Piero Straffa Cambridge Cambridge Univ Press, 1951

Rolph, Earl "Principles of Debt Management " *A.E.R* 47 (June 1957) 302–20

Reply to Feldstein and Buchanan

Robert J. Barro

University of Rochester

Martin Feldstein and James Buchanan have raised a number of critical points concerning my government-debt paper. For the reasons detailed below, I believe that my central thesis has not been impaired by these criticisms.

I. Feldstein

Feldstein's comment deals with my conclusion that government debt (and social security) will not add to net wealth in a model with operative intergenerational transfers (either from old to young or from young to old). He accepts this conclusion in the case of a static economy, but argues that it is incorrect when the economy is growing. It is important to distinguish two cases—the first in which the growth rate, g (the sum of population growth and growth in real income per head), is below the steady-state (real) rate of return, r, and the second where $g \geq r$.[1] The first case is straightforward and turns out to require no modifications to my previous results. Feldstein's contrary conclusion is the result of an error in calculating present values of future tax liabilities. The second case is both more interesting and more difficult, and I present some analysis of it below.

Suppose that the economy is in a steady state with output growing at rate g. The (real) rate of return is constant and equal to r. Consider, first, the case in which $r > g$. In this situation Feldstein (pp. 333–334) considers the impact of government debt that grows at rate g, so that the ratio of debt to income remains constant over time.[2] He argues, " some

---

[1] Feldstein also deals with a third case where the gross-of-tax rate of return is above g, but the net-of-tax rate is below g. Since the main purpose of this case is to make more plausible the situation where g is at least as large as the net rate of return, it is not necessary for me to deal with it separately.

[2] The limitation of the steady-state growth rate of the public debt to g can be rationalized by assuming that the value of the outstanding stock of debt at any point in time is bounded by the government's collateral, which I assume can be measured by the present value of future taxing capacity. If taxes are limited to a fraction, τ, of national income, y, then the present value, calculated at the current date t, of future taxing capacity is $\tau y(t)/(r - g) = \tau y(0)e^{gt}/(r - g)$. Hence, the government's collateral grows at rate g and is finite for any t if $r > g$. If government debt grew at a rate faster than g, then the outstanding stock would eventually exceed the government's collateral.

[*Journal of Political Economy*, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved

future taxes would have to be levied to limit the growth of the government debt. . . . For example, if . . . the debt grow[s] at the rate of growth of the economy, taxes will be required to finance only the fraction $[1 - g/r]$ of future interest payments" (p. 332). The last statement is correct but does not imply his conclusion that "only in the special case of a static economy ($g = 0$) is the present value of required future taxes equal to the present value of the debt itself" (p. 333). The fallacy is that the fraction, $(1 - g/r)$, applies to a base that is larger (increasingly over time) than it would have been if the debt had remained constant rather than growing at rate g.

Consider an initial debt issue of amount $B(0)$ and a path of subsequent issues that make the outstanding stock of debt grow at rate g. Hence, $B(t) = B(0)e^{gt}$, and the amount of debt finance at any date $t > 0$ is $dB/dt = gB(t)$. Taxes levied at date t are the amount needed to finance interest payments net of debt finance, $rB(t) - gB(t)$. The present value of these future taxes, discounted at rate r, is

$$\int_0^\infty (r - g)B(t)e^{-rt}\,dt = (r - g)B(0)\int_0^\infty e^{-(r-g)t}\,dt = B(0).$$

Hence, as would be expected, the present value of the future taxes coincides with the amount of the initial debt issue for any growth rate g (as long as $g < r$). Therefore, the presence of growth in the economy (at a rate below r) leaves unchanged the conclusion that government bonds are not net wealth.

Consider, now, the second case where $r \leq g$. (The situation with $r < g$ in a steady state implies inefficient capital overaccumulation, as discussed in Phelps [1966].) In this case it is possible to finance all of the future interest payments without incurring any future taxes by having the debt grow forever[3] at rate r.[4] In this situation, where $r \leq g$ in a steady state, it appears that debt issue would be regarded as net wealth and would therefore raise aggregate demand. Further, it seems that a sufficient amount of debt issue would cause enough of a shift from saving to consumption so that the steady-state rate of return would be raised to (just) exceed the growth rate.

---

[3] In one respect this possibility hinges on an infinite horizon—any finite truncation (corresponding, say, to a fixed date for the end of the world or to an exogenous liquidation date for the debt) would restore equality between the present value of taxes and the amount of initial debt. On the other hand, an individual's effective horizon for capitalizing taxes would also not be infinite if intergenerational transfers were not (expected to be) operative for some generations far in the future. In this sense it is the relative values of the horizons for the world and the individual that are important, rather than whether either is infinite. A further consideration, pointed out by Starrett (1970), is that the end of the world, if it is anticipated, is presumably stochastic. Starrett argues (p. 709) that a stochastic end of the world is equivalent to an infinite horizon with an appropriate specification of the discount rate applied to future utilities.

[4] The debt could also grow at rates above r to finance a continuing flow of transfer payments. The government's collateral, as calculated in n. 2 above, is now infinite.

The important issue is whether the economy would ever be in a steady state where $r \leq g$. The reason that government debt issue can move the economy out of this situation is that the government's collateral, which grows at rate g over an infinite horizon (see n 2), allows the debt to expand forever at rates at least equal to r Are there any private individuals or institutions that would have a similar capacity and incentive in the absence of government intervention? In the overlapping-generations models of Diamond (1965) and Cass and Yaari (1967), there are no economic units with a sufficiently long time perspective to engage in this type of activity, so that a competitive equilibrium with $r \leq g$ in a steady state is possible On the other hand, in a model of a utility-maximizing immortal family, as in Sidrauski (1967), the solution with $r \leq g$ in a steady state would be untenable In this situation the family would be motivated to play the same game as the government by issuing debts in exchange for current consumption, financing these debts by allowing them to grow at a rate at least equal to r but not higher than g—a rate that would not exceed the growth rate of the family's collateral—and thereby never reducing future consumption [5] Hence, a competitive equilibrium would have to be in the (efficient) region where $r > g$ in a steady state In a situation with finite-lived individuals, it seems that operative intergenerational transfers would play an equivalent role. In particular, when $r \leq g$ in a steady state, transfer payments from young to old (which I referred to as arising from a "gift motive" in my earlier paper) that grow at a rate at least equal to r but not higher than g would be feasible—in the sense that the ratio of transfers to income would not grow over time—and would raise at least the present consumption level without reducing any future levels of consumption [6] It seems that the operation of these types of transfers would insure that a competitive equilibrium is in the (efficient) region where $r > g$ in a steady state My main hesitancy about this conclusion is that I have been unable to demonstrate that the transfer scheme, which can feasibly raise the consumption level of all family members (living in finite time) if $r \leq g$, would be consistent with utility maximization as viewed by each family member separately (Hopefully, Brock and Scheinkman, well-known experts on the economics of infinity, will provide a final resolution of this question.) Accordingly, at this point I cannot rigorously rule out the existence of the case, $r \leq g$, in which debt issue would raise net wealth Nevertheless, I would be surprised if many advocates of standard fiscal policy would be happy to rest the argument for the efficacy of such policy on the possible existence of this case

---

[5] Thompson (1967, p 1205) claims that a corporate firm, which has an infinite legal life, would be motivated to carry out similar policies However, I cannot see why this policy is generally feasible, since the firm's collateral would not automatically grow at rate g; presumably, its net earnings are not a fixed fraction of national income.

[6] Reductions of bequests from old to young would have an equivalent effect, but it seems that bequests would not be operative in a steady state where $r \leq g$, at least if bequests are motivated solely by concern for the utility of descendants (and assuming preference for a unit of own consumption over that of a descendant, starting at equal values of consumption per head)

Finally, Feldstein observes near the end of his comment that the effect of government bond issue on future generations is complicated and is therefore likely to be neglected by current generations. It is plausible that computational difficulties would add to individual uncertainty, and the model could be usefully supplemented to make that uncertainty explicit However, it is much less clear that this complication would imply systematic errors in a direction such that public debt issue raises aggregate demand One could argue at the same level of rigor that government deficits make people sufficiently nervous to reduce their consumption demand when taxes are replaced by public debt issue As a general point it is difficult to identify a substitute for rational behavior as a basis for forming predictions in theoretical models

II. Buchanan

Buchanan begins his comment by pointing out my failure to acknowledge an intellectual debt to David Ricardo I readily accept this criticism, although it has no substantive bearing on my results

I should indicate, first, that my analysis was directed toward fiscal operations that involve shifts between tax and debt finance for a given volume of public expenditures Shifts in the expenditure level have real effects that depend on the degree of substitutability between public and private expenditures in individual utility functions (Bailey 1971, chap 9) and on the direct productivity of public expenditures (Grossman and Lucas 1974), and it did not seem necessary to become involved in these issues

Concerning a public debt issue that corresponds to either a transfer payment or a lump-sum tax cut, Buchanan states. " persons who purchase these bonds must draw down private investments or reduce private consumption " This assertion is incorrect since, as a matter of a budget constraint, the (representative) individual can use the full amount of his transfer payment (tax reduction) to just purchase the bonds, the issue is whether, in fact, individuals will be motivated to save the full amount of their increased disposable resources Buchanan then says, " . it is impossible that 'the increase in *B* [bonds] implies a one-to-one increase in the asset supply '" and later " the government bonds will replace private bonds or other assets " In fact, private debt issues are not replaced because of any direct budget-constraint requirements, but rather will be crowded out if the public debt issue leads to an excess supply of earning assets (at initial interest rates) and therefore to an increase in interest rates An excess supply results if individuals do not respond to the debt issue by saving (in the form of government bonds) the full amount of the transfer payment (tax cut). The one-to-one increase in saving (demand for bonds) arises if the future tax liabilities implied by the government debt are fully capitalized, so that perceived wealth is unchanged and individuals therefore maintain the same level of consumption demand Hence, Buchanan's statement that " . this neglects the possible effects of the sale of bonds on displaced private borrowers" does not make sense If government bonds are not perceived as net wealth, then the demand for bonds rises one-to-one with the supply, there is no change in interest rates, and no displaced private borrowers

The basic analysis can be extended to view government bonds and future tax liabilities as two elements in a larger household balance sheet with imperfect substitutability among the variety of assets and liabilities (because of differences in characteristics of risk and "liquidity," etc) In this context the demand for bonds would rise one-to-one with the supply (in the case of a "helicopter drop" of bonds) only if the bonds and the associated capitalized tax liabilities were perfect substitutes [7] More generally, one could consider the different risk and liquidity characteristics of bonds and the tax liabilities, and also bring in "imperfect" private capital markets (as I did in a preliminary way in Sections II–IV of my paper) The implications of these considerations for the net effect of public debt issue on aggregate demand hinge on such issues as whether the government or the private economy is more efficient at the margin in producing liquidity services or carrying out loans, and on how closely (uncertain) individual tax liabilities are correlated with individual incomes There is no presumption from this extended theoretical analysis that bond-financed tax reductions (transfer payments) would increase aggregate demand

Buchanan briefly discusses empirical evidence on whether future tax obligations are capitalized He cites Feldstein's (1974) study of social security as evidence to the contrary Feldstein's main conclusion (p 920) is that the " effect of social security is to reduce personal saving by half of what it otherwise would be " In fact, Feldstein's results (p 917) show a significant effect of a "social security wealth variable" (which seems to measure primarily changes in the fraction of workers covered by social security and secondarily shifts in age structure) on consumption only when the years 1929–40 are added to the 1947–71 sample, suggesting that the variable may be operating principally as a dummy for the pre-1937 years for which it takes on a zero value Further, the effect of the social security variable over the entire sample (1929–40, 1947–71) is insignificant when an unemployment variable is included.[8] Accordingly, Feldstein's finding that social security has drastically reduced saving and capital accumulation must be regarded as highly tentative.

---

[7] In the sense that individuals were concerned only with the difference between the amount of government bonds and the amount of capitalized tax liabilities in their portfolios

Some support for the view that future taxes are capitalized is presented in empirical studies of consumption and saving behavior by Tanner (1970), Kochin (1974), and David and Scadding (1974)

Finally, Buchanan (p 340) adds the weight of his own casual observation to the empirical issue " the behavior of legislators seems to offer indirect evidence against the capitalization hypothesis Can anyone in the post-Keynesian world of 1975 seriously question the proclivity of politicians to expand public debt in preference to tax increases?" This hypothesis may be a bit loose to submit to empirical testing, but it is hard to reconcile it, in a general way, with the post-World War II behavior of the federal government The quantity of U S government debt that is privately held (excluding holdings of the Federal Reserve, social security trust fund, etc) was[9] $204 billion at the end of 1947 and $230 billion at the end of 1970—implying an average annual growth rate of 0 5 percent per year On the other hand, U S government tax receipts (including other receipts like social security payments) rose from $43 billion in 1947 to $192 billion in 1970—an average annual growth rate of 6 5 percent per year (If federal politicians had a "proclivity to expand public debt in preference to tax increases" over this period, it was more than matched by a "declivity" of the public) Despite the modest growth of privately held public debt, the ratio of federal expenditures to GNP rose from 13 percent in 1947 to 21 percent in 1970 Hence, the ready availability of debt finance is apparently not a necessary accompaniment to increases in the share of output absorbed by the federal government Since 1970 there has been a faster growth rate of public debt—to $271 billion at the end of 1974, implying a 4 1 percent average annual growth rate from 1970 to 1974 Over this period, taxes rose to $291 billion, implying a 10 4 percent average annual growth rate of taxes from 1970 to 1974 Hence, even in the most recent period, taxes grew faster than the stock of privately held federal government debt Further, the fraction of GNP accounted for by federal spending was almost constant from 1970 to 1974, suggesting that an acceleration of public debt is not a sufficient indicator of a rising share of government spending in total output.[10]

---

[8] For further discussion of these results, see Upton (1975)

[9] All data are from Council of Economic Advisers, *Economic Report of the President*, 1975 and earlier issues

[10] This note is a convenient place to correct some typographical errors in my original paper The date references to Patinkin's book on pp 1096, 1116, and 1117 should be 1965 On p 1099, the A_2^0 in equation (3) should be A_1^0 The A_1^0's that appear two and eight lines below this equation should be A_i^0

References

Bailey, M. *National Income and the Price Level.* 2d ed. New York: McGraw-Hill, 1971

Cass, D, and Yaari, M. "Individual Saving, Aggregate Capital Accumulation, and Efficient Growth" In *Essays on the Theory of Optimal Economic Growth,* edited by K Shell Cambridge, Mass.. M I.T. Press, 1967.

David, P, and Scadding, J. "Private Savings. Ultrarationality, Aggregation, and 'Dennison's Law.'" *J.P E.* 82, no 2 (March/April 1974)· 225–49.

Diamond, P "National Debt in a Neoclassical Growth Model" *A.E R.* 55 (December 1965) 1126–50

Feldstein, M "Social Security, Induced Retirement, and Aggregate Capital Accumulation" *J P E.* 82, no 5 (September/October 1974). 905–26.

Grossman, H, and Lucas, R F. "The Macroeconomic Effects of Productive Public Expenditures" *Manchester School* 42 (June 1974) 162–70

Kochin, L "Are Future Taxes Anticipated by Consumers" *J. Money, Credit and Banking* 6 (August 1974) 385–94

Phelps, E "An Essay in Dynamical Efficiency" In *Golden Rules of Economic Growth.* New York Norton, 1966

Sidrauski, M "Rational Choice and Patterns of Growth in a Monetary Economy" *A E R Proc* 57 (May 1967) 534–44

Starrett, D "The Efficiency of Competitive Programs" *Econometrica* 38 (September 1970) 704–11

Tanner, J E "Empirical Evidence on the Short-Run Real Balance Effect in Canada" *J Money, Credit and Banking* 2 (November 1970)· 473–85

Thompson, E. "Debt Instruments in Macroeconomic and Capital Theory" *A E R* 67 (December 1967) 1196–1210

Upton, C Review of A Munnell, *The Effect of Social Security on Personal Saving* *J P E* 83, no 5 (October 1975) 1090–92

Earnings and Individual Variations in Postschool Human Investment

Charles B. Knapp
University of Texas at Austin

W. Lee Hansen
University of Wisconsin—Madison

This paper examines the effect on earnings of variations among individuals in postschool human investment. Several studies have recently employed the human-capital model initially posited by Becker and Chiswick (1966) to investigate the earnings-generation process.[1] These studies have been hampered by the absence of detailed information on the postschool investment patterns of individuals. As a result, they have been forced to assign individuals of the same age and education identical amounts of postschool investment.[2] Since postschool investment presumably varies with other factors as well, the explanatory power of the model is thereby limited, and, under certain conditions, biased estimates of the model's parameters will result. This paper reformulates the Becker-Chiswick model to allow for additional variations among individuals in the amounts of their postschool investment.

I. Individual Variations in the Postschool Investment Model

The Becker-Chiswick model may be summarized in its simplest form as

$$ln\,y(t) = A + R_s S + R(t)P(t) + ln[1 - k(t)] + u(t)_i, \qquad (1)$$

where

t = years of work experience (assume continuous and equal to age − years of schooling − 5);
$ln\,y(t)$ = the natural logarithm of net earnings (potential − foregone earnings) in year t,

---

The authors wish to thank Richard S. Higgins, Allan G. King, and Jacob Mincer for helpful comments on earlier drafts of this paper. The authors are responsible for remaining deficiencies. The support of the University of Wisconsin Graduate School and the Institute for Research on Poverty are gratefully acknowledged.

[1] See particularly Becker and Chiswick (1966), Johnson (1970), and Mincer (1974).

[2] An exception to this practice is Mincer and Polachek (1974) who explicitly consider the impact of labor-force participation.

[*Journal of Political Economy*, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved.

A = the natural logarithm of the earnings that would be obtained without human investments;
R_s = the average rate of return to schooling,
S = years of schooling,
$R(t)$ = the average rate of return to postschool investment (assume constant over t),
$k(t)$ = $C(t)/E(t) = 1 - Y(t)/E(t)$, the ratio of investment costs to potential earnings in year t,
$P(t)$ = $\int_0^t k(t)dt$, the number of "equivalent years" previously devoted to postschool investment,
$U(t)_1$ = a random disturbance term with mean = 0 and constant variance (for eq [1])

In (1) the level of postschool investment in year t is measured by the investment fraction $k(t)$ The manner in which $k(t)$ varies over the working life must be assumed since it is not observed directly [3] The most widely employed specification has been that $k(t)$ declines linearly from some initial investment fraction, $k(0)$, to zero at the end of the postschool investment span [4] Thus $k(t) = k(0) - \{t[k(0)/T]\}$, where T is the length of the investment span [5] For a given value of T, the postschool investment pattern $[k(t), t = 0, 1, \quad, T]$ is completely specified by the choice of $k(0)$

Empirical estimation of (1) has generally relied on cross-section data Because $k(0)$ is not observed for individuals in a cross section, (1) is typically written as

$$ln\, y(t) = A + R_sS + R(t)k(0)t - R(t)k(0)\frac{t^2}{2T} + ln\left[1 - k(0) + \frac{tk(0)}{T}\right] + u(t)_2, \tag{2}$$

where $k(0)$ is an estimated parameter In (2), postschool investment varies only with experience, and, since experience is assumed to equal age minus years of schooling minus five, individuals of similar age and education are assigned identical quantities of postschool investment Although this specification has been employed primarily for empirical convenience, it detracts, at least on a theoretical level, from the model's explanatory power Relative variations in earnings attributable to variations in postschool investment patterns within age-education groups are not accounted for

---

[3] Haley (1973) derives a theoretical specification for $k(t)$ assuming individuals attempt to maximize lifetime disposable income

[4] See, for example, Johnson (1970), Malkiel and Malkiel (1973), Mincer (1974), and Mincer and Polachek (1974)

[5] In this paper we assume $T = 20$ It can be established from (1) that, ignoring the disturbance term, T equals the experience level at which earnings peak The assumption that $T = 20$ for annual earnings has also been employed by Mincer.

Even with longitudinal data it is still not possible to observe $k(0)$ directly for individuals. However, in this case an approximation can be made. Since $k(0) = 1 - Y(0)/E(0)$, and $Y(0)$ is observed, $k(0)$ can be approximated if an estimate of $E(0)$ is available. If annual dollar investment costs are constant over the relevant range of the working life, $E(0) = Y(\hat{t})$, where $\hat{t} = 1/R(t)$.[6] By using Johnson's (1970, p. 558) estimates of $R(t)$ between socioeconomic groups, $E(0)$ and hence $k(0)$ can be estimated for each individual.[7] To the extent that annual dollar investment costs decline, the estimate of $k(0)$ will be biased downward.

Let $k(0)^i$ be such an approximation of the initial investment fraction for the ith individual ($i = 1, \quad , n$). Equation (2) can then be written for individual i as

$$ln\, Y(t)^i = A + R_s S^i + R(t)[k(0)^i]\left(t - \frac{t^2}{2T}\right) + ln\left[1 - k(0)^i + \frac{t}{T}k(0)^i\right] \\ + U(t)_3 \qquad (3)$$

Equation (3) is estimable only with nonlinear regression techniques. An estimable variant can, however, be obtained by transferring the logarithmic term to the left-hand side. Since $Y(t)^i = [1 - k(t)^i]E(t)^i$, the dependent variable in this transformed equation is the logarithm of potential earnings. Thus

$$ln\, E(t)^i = A + R_s S^i + R(t)[k(0)^i]\left(t - \frac{t^2}{2T}\right) + U(t)_4 \qquad (4)$$

Let $k(0)^i = \bar{k}(0) + k(0)^{i*}$, where $\bar{k}(0)$ is the mean of $k(0)^i$ and $k(0)^{i*}$ is the deviation from that mean. Equation (4) can then be written more generally as

$$ln\, E(t)^i = A + R_s S^i + R(t)^E\bar{k}(0)\left(t - \frac{t^2}{2T}\right) + R(t)^D k(0)^{i*}\left(t - \frac{t^2}{2T}\right) \\ + U(t)_5 \qquad (5)$$

Equations (4) and (5) allow for variations within age-education groups in postschool investment. In (5) the variations are characterized by $k(0)^{i*}$, which in turn uniquely specifies individual postschool investment patterns. The parameter $R(t)^E$ can be interpreted, as it is in specifications such as (2) which do not allow variations in $k(0)^i$, as the average gross rate of return to experience alone The parameter $R(t)^D$, by measuring the effect of $k(0)^{i*}$, captures the impact of especially intensive (or disintensive) human investment while holding experience constant. The advantages of (5) over a formulation which does not account for variations within age-education groups in postschool investment can be illustrated by viewing (6)

---

[6] Experience level $\hat{t}$ in this case is the "overtaking" point, that is, the experience level where actual earnings are equal to the earnings which would have been attained with no postschool investment (see n. 8).

[7] Johnson's rates are actually estimated for marginal investment decisions—for example, the decision once high school is completed to graduate from college and undertake subsequent postschool investment. But given final educational attainment, the estimated rates are so similar (differing at most by two-tenths of a percent) that the choice of rates is inconsequential. Although this method of approximating $k(0)$ assigns the same value of $\hat{t}$ to all individuals in a socioeconomic group, the final estimate of $k(0)$ is unique, since $Y(\hat{t})$ is observed for each individual.

$$\ln E(t)^i = A + R_s S^i + R(t)^E \bar{k}(0)\left(t - \frac{t^2}{2T}\right) + U(t)_6. \qquad (6)$$

Equation (6), when considered relative to (5), is subject to specification error, namely, omission of the relevant explanatory variable $k(0)^{i*}$ $[t - (t^2/2T)]$

One advantage of (5) over (6) is that inclusion of the term $k(0)^{i*}$ $[t - (t^2/2T)]$ guarantees an increase in the model's explanatory power, as measured by the coefficient of determination (R^2) From an interpretive standpoint, this increase results because (6), since it does not recognize variations in postschool investment within race-education groups, underestimates the percentage of relative earnings variation attributable to human investments Mincer (1974) has argued, using data on the relative variation in actual earnings in the "overtaking" set,[8] that if full information were available on postschool investments for individuals, the resulting R^2s would more than double

There is also reason to believe that the estimation of (6) rather than (5) will lead to biased estimates of the parameters A, R_s, and $R(t)^E$ Kmenta (1971, pp 392–95) provides a method of calculating the parameter biases which result from a misspecification such as (6) Consider (7), which expresses for (6) the omitted variable as a linear function of the included explanatory variables

$$k(0)^{i*}\left(t - \frac{t^2}{2T}\right) = \beta_0 + \beta_1 S^i + \beta_2[\bar{k}(0)]\left(t - \frac{t^2}{2T}\right) + U(t)_7 \quad (7)$$

Let b_0, b_1, and b_2 be the ordinary least-squares estimates resulting from (7) and $\hat{A}$, $\hat{R}_s$, and $\hat{R}(t)^E$ be the least-squares estimates resulting from (6) The parameter biases are then given as

$$\begin{aligned} E[\hat{A}] &= A + R(t)^D b_0 \\ E[\hat{R}_s] &= R_s + R(t)^D b_1 \\ E[\hat{R}(t)^E] &= R(t)^E + R(t)^D b_2 \end{aligned} \qquad (8)$$

If $R(t)^D$ is positive, the direction and extent of the biases will depend on b_0, b_1, and b_2

---

[8] Since relative earnings variation at the "overtaking" point is due to factors other than postschool investment (see n 6), these results can be employed indirectly to calculate the relative earnings variation due to postschool investment. The Mincer sample includes only adult, white, urban men in the United States in 1959.

TABLE 1

PREDICTIONS OF BIASES RESULTING FROM THE ESTIMATION OF (7)

Dependent Variable	Constant (b_0)	Schooling (b_1)	$\bar{k}(0)[t - (t^2/2T)]$ (b_2)	R^2	N
$k(0)^{t*}\left(t - \frac{t^2}{2T}\right)$	1 3343	− 1171** (0119)[a]	1230** (0342)	0250	5,100

SOURCE —Johns Hopkins Life History Study
[a] Standard errors are in parentheses
* Significant at the 05 level
** Significant at the 01 level

II. Empirical Results

The data employed in this study are from the Johns Hopkins Retrospective Life History Survey [9] The data describe individual occupational, educational, familial, residential, and earnings experiences from age 14 to the time of interview The population is comprised of males aged 30–39 in 1968 residing in households in the United States A sample of 324 observations is analyzed here The number of pooled observations (each of the 324 individuals at various experience levels) is 5,100

Table 1 presents a regression estimate of (7) The results indicate that the base earnings capacity ($ln^{-1}A$) will be biased up, the rate of return to schooling (R_s) will be biased down, and the rate of return to experience $[R(t)^E]$ will be biased up

Equation (A) of table 2 is a regression estimate of (5), and (B) is a regression estimate of (6) [10] The empirical results are consistent with the hypotheses outlined above When the additional variations in postschool investment are accounted for, the coefficient of determination increases by almost 50 percent, from R^2 = 0644 in (B) to R^2 = 0951 in (A) Although this is smaller than the increase predicted by Mincer, this is to be anticipated, as (5) captures some but certainly not all of the variation in postschool investment within race-education groups It should also be noted that the magnitudes of these coefficients of determination are not directly comparable to the results of, for example, Mincer (1974,

---

[9] This survey was initiated by James S Coleman and Peter H Rossi

[10] We assume the error terms in (5) and (6) may be serially correlated for the same individual but are not dependent for different individuals Ordinary least-squares estimates of (5) and (6) will thus be unbiased and consistent, although not efficient But with such a large sample efficiency is not a central concern, and, as we shall see, all estimated coefficients in table 2 are significant at the .01 level anyway We have consequently estimated (5) and (6) by ordinary least squares If the error term were not independent for different individuals, an alternative estimation technique such as error components would be required (Wallace and Hussain 1969)

TABLE 2

INDIVIDUAL VARIATIONS IN POSTSCHOOL INVESTMENT

Equation	Dependent Variable	Constant	Schooling	$\bar{k}(0)[t - (t^2/2T)]$	$k(0)'^{*}[t - (t^2/2T)]$	R^2	N
(A) .	$ln\,E(t)'$	7 7676	0502** (0029)	1077** (0082)	0052** (0004)	0951	5,100
(B)	$ln\,E(t)'$	7 8235	.0454** (0029)	1133** (0024)	*	0644	5,100

SOURCE.—Johns Hopkins Life History Study
*Not applicable
*Significant at the 05 level
**Significant at the 01 level

p. 92), since we are employing potential rather than actual earnings as the dependent variable

We also find that the estimation of (5) in comparison with (6) results in a lower estimate of the base earnings capacity $(ln^{-1}A)$,[11] a higher estimate of the rate of return to schooling, and a lower estimate of the rate of return to experience $[R(t)^E]$ The directions of these differences are consistent with the predictions of bias given in table 1 The most important of these results are those which deal with education and postschool investment If the results can be generalized, they suggest that previous efforts at estimating the Becker-Chiswick model using aggregate data which obscured variations in postschool investment within race-education groups may have understated the advantages of education and overstated the advantages of labor-force experience [12]

Also of interest is the estimate given in (A) of table 2 for $R(t)^D$, the average rate of return to deviations from the mean number of "time equivalents" spent in postschool investment The estimated return (0052) is approximately one-twentieth of the estimated return to experience alone (1077) This large difference suggests that earnings are relatively insensitive to deviations from the average amount of postschool human investment That is, an individual gains little by choosing a $k(0)^i$ above the norm or, conversely, loses little by choosing a $k(0)^i$ below the norm This may reflect the tendency of employers to reward, through such practices as seniority, the nominal amount of experience a worker has rather than any especially intensive human investment

III. Summary and Conclusions

This paper has refined the widely employed Becker-Chiswick human-capital model to allow for variations within age-education groups in postschool human investment The percentage of relative (potential) earnings variation explained rises by about 50 percent when these variations are accounted for There is also evidence that not allowing for these variations will lead to understating merits of the rewards to education and to overstating merits of the rewards to experience We additionally find that the returns to above-average postschool investment are quite small

We conclude, as was earlier suggested by Mincer (1974), that the explanatory power of human-capital theory may be greater than earlier empirical work has demonstrated One might anticipate that as more sophisticated longitudinal data become available (for example, data specifying the experience variable more precisely) further increases in the model's explanatory power will be realized Moreover, the development of such data sources may allow us to regard the estimates of the parameters of the human-capital model with more certainty

---

[11] For (5) ln^{-1} (7 7676) = 2362 77, and for (6) ln^{-1} (7 8235) = 2498 64 These base earnings capacities include the impact on the cohort of secular income growth, assuming uch growth in a given calendar year is independent of experience levels (Johnson and Hebein [1974, p 605])

[12] It is, of course, possible that the techniques employed in this paper have introduced other specification errors For example, $k(0)i$ may be measured with error (see nn. 6 and 7)

References

Becker, G , and Chiswick, B "Education and the Distribution of Earnings" *A E.R* 56 (May 1966) · 350–69

Haley, W "Human Capital The Choice between Investment and Income." *A E R* 63 (December 1973) 929–44

Johnson, T "Returns from Investment in Human Capital" *A E R* 60 (September 1970) 546–60

Johnson, T., and Hebein, F "Investment in Human Capital and Growth in Personal Income 1956–65" *A E R* 64 (September 1974) 604–15.

Kmenta, J *Elements of Econometrics* New York Macmillan, 1971

Malkiel, B , and Malkiel, T "Male-Female Pay Differentials in Professional Employment" *A E R* 63 (September 1973) 693–705

Mincer, J *Schooling, Experience, and Earnings* New York Nat. Bur. Econ Res , 1974

Mincer, J , and Polachek, S "Family Investment in Human Capital Earnings of Women" *J P E* 82, no 2, suppl (March/April 1974) S76–S110

Wallace, T , and Hussain, A "The Use of Error Components Models in Combining Cross-Section and Time Series Data" *Econometrica* 37 (June 1969) 55–72

The Conundrum of the Low Issue of National Bank Notes

John A James

University of Virginia

In the late nineteenth century the United States experienced a prolonged period of deflation, with the Warren-Pearson wholesale price index declining from 136 in 1872 to 68 in 1896 (Warren and Pearson 1933, pp 12–13) Between 1879 and 1897 prices fell at a rate of 1 percent per year while the money stock increased at a 6 percent annual rate, which was markedly slower than the rate of increase after 1897. This relatively slow rate of growth of the money supply resulting in the deflation has frequently been attributed to the scarcity of gold over the period. To be sure, increases in the stock of high-powered money did account for two-thirds of the monetary growth which occurred between 1879 and 1897 (Friedman and Schwartz 1963, p 121), but the scarcity argument is somewhat incomplete since gold was not the only constraint on the growth of the stock of high-powered money The volume of high-powered money at the time was nowhere near the maximum allowable amount

National bank notes also were a component of high-powered money. The National Banking Acts of 1863 and 1864 authorized national banks to issue notes after certain types of U S bonds were deposited with the treasury as collateral The amendments of 1882 limited note issue by individual national banks to 90 percent of the paid-in capital of the bank and 90 percent of the par or market value, which ever was lower, of the bonds used as collateral The Gold Standard Act of 1900 eased these requirements to allow issues up to 100 percent of the paid-in capital and the par or market value of the bonds [1]

Between the passage of the Resumption Act of 1875 and 1900 never more than 30 percent of the maximum allowable issue of notes was actually issued Some constraint other than just gold affected the growth of the stock of high-powered money What could have limited the issue of national bank notes? In fact, the problem is all the more curious in

---

The insights of Michael Lichstein are gratefully acknowledged

[1] For more detail about the legal provisions of note issue see, e g , *Report of the Monetary Commission of the Indianapolis Convention* (1898), chap on "History of the National Banking System "

Journal of Political Economy, 1976, vol 84, no 2]
© 1976 by The University of Chicago All rights reserved

view of Cagan's calculations of the profitability of note issue, which show it on the whole to have been a quite lucrative activity.[2] In only the period 1884–91 was the profit from note issue below the average return on national bank assets; between 1892 and 1897 the excess profit on note issue averaged 3.3 percent per year above other bank assets; after the liberalized requirements went into effect in 1900 the return averaged 25 percent per year by Cagan's calculations (Cagan 1965, pp. 88, 93).

The percentage of the maximum allowable amount of notes actually issued did fluctuate with the profitability of note issue, but in view of their apparent profitability why was the total number of notes issued so small? In the 1890s for example, note issue was quite profitable, but by 1900 only 28 percent of the maximum allowable had been issued. Friedman and Schwartz conclude, "We have no explanation for this puzzle," and allude to either obtuseness on the part of bankers in the face of a profitable opportunity or some unseen costs of note issue (1963, pp. 23–24).

I. The Goodhart Explanation

One attempt at such an explanation making explicit the unseen costs was by C. A. E. Goodhart (1965), who tried to explain the low issue of notes in the period 1900–1913 when note issue was quite profitable and the quantity of notes in circulation was rising rapidly. He argued that banks were reluctant to issue notes because "there was always a possibility, and after 1907 a probability, that the circulation privileges would be terminated. If this happened, the bonds held as collateral would fall in value" (p. 516). Over the period which Goodhart examines, 1900–1913, national bank notes outstanding more than doubled from $323.5 million to $733.7 million; in 1900 note issue amounted to 28 percent of the possible maximum and rose to 80 percent by World War I (Friedman and Schwartz 1963, p. 23). This risk explanation is designed to explain the retardation in the rate of increase of note issue after 1907 in view of substantial increases in its apparent profitability over that period. At best, this explanation is limited to the period 1907–13 and has little relevance to the question of low note issue in the 1890s; moreover it is not an operational hypothesis.

If note issue was so risky after 1908, then why did the proportion of

---

[2] Cagan assumes the notes obtained from the purchase of a U.S. bond could be used to purchase more bonds for note issue, and then the process is repeated. For simplicity we may assume the bank paid for the initial bond purchase with the notes obtained from its deposit plus enough of its own capital to make up the difference. The annual rate of return on invested capital from note issue up to 1900 then, is $(Pi/100 - E)/(P - 90)$, where P is the market price of the bond, i is its yield, E is the annual expense of issuing $90 in notes. This assumes the market price exceeded par value, which was true throughout the period (Cagan 1965, p. 87).

TABLE 1

RATIO OF OUTSTANDING CIRCULATION TO CAPITAL (%)

Date	New York City	Central-Reserve Cities	Other Reserve Cities	Country Banks
September 6, 1904	36.9	38.0	51.4	59.1
August 25, 1905	50.3	49.5	55.2	62.8
September 4, 1906	43.2	47.8	59.6	66.5
August 22, 1907	44.2	46.7	58.2	66.8
September 23, 1908	49.1	51.6	64.9	71.6
September 1, 1909	46.1	50.8	67.8	75.9
September 1, 1910	38.4	44.8	64.3	75.6

SOURCE.—Bell 1912, p. 56.

notes outstanding remain so high? Goodhart's answer lies in differing behavior of country and city banks. Between 1908 and 1913 the ratio of notes outstanding to paid-in capital declined from 51.6 to 42.9 percent and from 65.0 to 61.8 percent for central-reserve city and reserve city banks, respectively, while it rose from 71.6 to 79.1 percent for country banks (1965, p. 520). Thus, the increased note issue of country banks concealed the decline by city banks. Goodhart (p. 522) explains this relative decline in note issue by city banks in terms of lower costs of note issue for country banks. The marginal costs of note issue differed between city and country banks because of differences in redemption costs. Country banks could get rid of unwanted notes by depositing them in city correspondent banks and earning interest on them. On the other hand, city banks were forced to submit these notes for redemption in slack periods when they could not keep them in circulation, since they did not constitute lawful money reserves. During the transit period after sending the notes off to be redeemed and before receiving lawful money back, the redeeming bank earned no interest at all on them.

This interim period between the shipping of the notes and receiving the redemption money was usually about a week or 2 (Goodhart 1965, p. 521). If the loan rate in reserve cities during this period was 6 percent, which is a substantial overestimate for most reserve cities, the opportunity costs of redemption in foregone earnings amounted to 0.25 percent of the value of the notes redeemed for a 2-week redemption process. One-quarter of 1 percent cost differential is a trifling one to explain the change in country and city bank issue of notes.

Moreover, consider table 1. If the cost-differential argument were valid, it should hold in every period, since country banks always had the option of sending their notes to city correspondents. In table 1 we see that between 1904 and 1910 central-reserve city and reserve city banks as well as country banks expanded their note issue. A more plausible explanation for the relative increase in country bank note issue and a more general explanation for the low issue of national bank notes will now be offered

II. An Opportunity Cost Explanation

This "puzzle" of the low issue of national bank notes becomes clear as soon as the process of bank portfolio selection is considered A bank has to decide the relative amounts of notes and deposits to hold in its liability portfolio when making loans If the rate of return on loans is very high, as was the case in the South and West, then clearly it is going to be more profitable to make loans through creating deposits rather than to buy government bonds and then issue notes (in the form of loans) on 90 percent of par value

Lichstein (1968) has argued that Cagan's profitability calculations break down in practical application Cagan assumed that notes issued on a bond were used to help pay for the bond If bond purchases were made through a broker, usually at some distance from the bank, it is unlikely that such transactions were made by notes If they were, the broker would undoubtedly have deposited them with another bank, which probably would have redeemed them Cagan assumed the bank could keep all its notes outstanding in the process of purchasing more bonds

Instead the bank is faced with two basic choices of what to do with an increment in deposits It can loan out the funds directly or purchase bonds with them and loan out notes equal to 90 percent of their par value At high interest rates it will be more profitable to loan out the total directly since otherwise it would be losing the high interest on the difference between the market price and 90 percent of the par value of the bonds Assume bonds sell at par and their yield is 2 5 percent We also assume the costs of note issue to be 1 percent, a plausible figure (Lichstein 1968, p 9) and, for computational convenience, that 85 percent of the par value of bonds is issued as notes Now the local interest rate X at which the return on note issue equals the return on loaning funds directly may be solved for

$$\begin{aligned} 2.5\% + 0.85X - 1\% &= X \\ X &= 10\% \end{aligned} \qquad (1)$$

At a loan rate of 10 percent the bank is indifferent between issuing notes or directly making loans, at rates above 10 percent the bank will find lending directly more profitable and issue no notes, ignoring diversification motives

So western and southern banks would issue very few notes because the opportunity costs of note issue, local loan rates, are so high. On the ther hand, in the East where interest rates were lower note issue could e very profitable, and we would expect to see a relatively large amount notes issued This simple idea formed the basis of the explanation of w note issue by the Indianapolis Monetary Commission, as summarized Taylor (1898, p 320): "Whether or not a given rate is to be accounted w depends upon the alternative investments which are open to the nkers In the older districts such alternative investments at times bring ven lower returns than government bonds [In the West] banks ill find opportunities for use of all their resources at six to eight percent aturally enough they will invest the minimum possible in government onds In consequence, those portions of the country which have the reatest need for notes have as a rule the smallest supply "

The commission was misled by an erroneous calculation of the profitability of note issue and thus compared local loan rates with the fixed eturns on government bonds but nevertheless recognized the geographic ariation in the pattern of note issue The percentage of maximum permissible note issue that was actually issued by banks, which was taken to ave been determined by the provision that national banks could issue otes up to 90 percent of the value of their capital, in 1897 in such Eastern tates as New Hampshire and Rhode Island was 55 9 and 53 9, respectively, while in North Carolina it was 3 2, in Mississippi, 4 2, in Arkansas, 0 8, in Nebraska, 2 6, in Kansas, 6 9[3] The reason for the low note issue s clear A large part of the country, the West and South, issued virtually o notes at all because banks had more profitable alternative uses for their funds

This explanation is also supported by table 2, compiled by Charles Dunbar (1904) who also emphasized the geographic variation in the rate of note issue although he, too, was misled by an erroneous calculation of the profitability of note issue. In 1896 in Texas, 88 national banks had capital of the $50,000 minimum requirement and 86 of these held the exact minimum required amount of government bonds, of the 76 Texas national banks with capitals between $50,000 and $100,000, 63 held the exact minimum of government bonds It is precisely in these rural areas that loan rates were the highest Similarly, in 1896 in Nebraska 113 national banks, having an aggregate capital of $10,975,000, held only $368,650 of government bonds in excess of the minimum requirement, and $333,000 of the excess was held by banks in the cities of Lincoln and Omaha (Dunbar 1904, p 233)

One of the ironies of this geographical difference in note issue is that he demand for currency against demand deposits is generally thought o have been higher in the West and South However, it was in those

---

[3] These percentages represent the discretionary note issue of banks, excluding that on S bonds required to be held as part of bank capital, i e , a percentage of the amount at may be issued in excess of the required bond deposit (*Report of the Monetary Commission the Indianapolis Convention* 1898, p 192)

TABLE 2

BONDS HELD FOR CIRCULATION IN EXCESS OF MINIMUM REQUIREMENT, OCTOBER 1896 ($ MILLION)

Region	Bonds Held for Circulation	Minimum Required	Percentage of Excess
New England	65.4	21.6	67
Middle States	87.9	28.9	67
North Central	45.6	24.9	45
South and Southwest	21.3	14.2	33
Missouri River	9.2	7.1	23
Central and Mountain	3.9	2.9	29
Pacific	3.9	3.0	24

SOURCE.—Dunbar 1904, p. 233.

rural western and southern states, where local loan rates were high, that bank notes were hardly issued.

The low issue of national bank notes can be explained by the high opportunity costs of note issue in the form of local loan rates in southern and western states.[4] What then caused the substantial increase in note issue by country banks and, to a lesser extent, by other reserve city banks as partially evidenced in table 1? The solution to a large extent lies in the declines in local interest rates over the period. Especially in interior areas, where most of the country banks were located, local interest rates fell more rapidly over the last part of the nineteenth century, narrowing the differentials with eastern rates.[5] As the local loan rates of country banks declined, it eventually became profitable for those banks to issue notes. Similarly, by the turn of the century a substantial number of reserve cities were fairly small ones in the West and Midwest; as loan rates fell there also, it became profitable to issue notes. In contrast, in central-reserve cities, which were well developed and stable money markets, the percentage of notes issued changed relatively little.

This hypothesis may be tested empirically. To be sure, this model of the determinants of note issue is a very simple one in which only the supply side is considered.[6] The percentage of the maximum allowable notes issued depends only on local opportunity costs. The relative difficulty of keeping notes outstanding is not considered explicitly, since data on demand influences such as the currency-deposit preferences of the local community are not available [7] The ***Report of the Monetary Commission*** (1898, p 375) also argued that note issue was smaller in western regions because of the riskiness involved in the high seasonal variability of demand in agricultural regions

---

[4] Cross-section regressions for selected time periods indicate that the local interest rate, the opportunity cost of note issue, is a significant factor in accounting for geographical variation in note issue. The higher the local interest rate, the smaller the percentage of note issue.

[5] For more information about declining interest rates and the narrowing of interregional interest rate differentials see Davis (1965) and James (1974, chap. 1).

[6] It is implicitly assumed that some degree of interregional capital immobility existed, so that profitability differentials were not arbitraged over time by interregional flows of funds. Despite the existence of such mechanisms as interbank borrowing and lending, a perfect capital market did not exist in the postbellum United States (James 1974, pp. 477–86).

Does variation in local interest rates geographically and over time explain differences in percentage of notes issued? Cross-section time series over the period 1888–1911 were estimated for both country banks by state and reserve city banks using a two-round BAN (Best Asymptotic Normal) estimator in which the error term is considered as the sum of a group specific and an individual component (Balestra and Nerlove, 1966) The dependent variable, PCXCS (Percentage of Excess Bonds Held), is the percentage of excess bonds held for circulation in relation to minimum required bond holdings The principal explanatory variable, R, is the difference between the local loan rate[8] and X in equation (1), the interest rate at which the bank is indifferent between issuing notes and creating deposits, ignoring diversification motives [9] In other words, R measures the relative profitability of notes versus deposits, the more positive R is, the more profitable it is for the local bank to create deposits, SEAS is a semiannual seasonal dummy The results for country banks, with t-statistics in parentheses, appear below

$$\begin{aligned} \text{PCXCS} &= \underset{(-46\ 26)}{-5\ 125\,R} - \underset{(-17\ 35)}{50\ 450\ \text{SEAS}} + \underset{(96\ 75)}{71\ 821} \qquad (2) \\ R^2 &= 595 \qquad F(2{,}1792) = 1314\ 6 \end{aligned}$$

Similarly, the results for reserve city banks are

$$\begin{aligned} \text{PCXCS} &= \underset{(-18\ 99)}{-2\ 623\,R} + \underset{(0\ 96)}{4\ 087\ \text{SEAS}} + \underset{(29\ 99)}{63\ 594} \qquad (3) \\ R^2 &= 327 \qquad F(2{,}738) = 179\ 6 \end{aligned}$$

---

[7] To the extent that currency-deposit preferences of an area are stable over time, this should be taken care of in the group effect of the cross-section time series Moreover, if currency demand is higher in rural areas, *ceteris paribus* leading to an increase in note issue, and local loan rates are high there also, then omission of the demand-side influence biases the results against the hypothesis

[8] Local interest rates were computed from the balance sheets and earnings statements of national banks reported by the Comptroller of the Currency, the series are semiannual from 1888 to 1911 (see James 1974, chap 6A)

[9] The effect of the Gold Standard Act of 1900 in raising the maximum note issue from 90 to 100 percent of the par or market value, whichever was lower, of the U S bond purchased may be incorporated directly into variable X, since through eq (1) it is equivalent to an increase in the interest rate at which the bank is indifferent between issuing notes or creating deposits However, the simple form of eq (1) has no solution when notes may be issued up to 100 percent of the value of the bonds because of the assumption that bonds sell at par In fact, over the period government bonds sold above par, so that the bank could loan out directly the total market value of the bond or could issue notes only on the par value Thus differential between the market and par value of government bonds was taken into account in computing the X series for the estimated equation

In both cases the effects of local interest rates are highly significant, with a 1 percentage point decline in local interest rates leading to an increase of more than 5 percentage points in the proportion of notes issued relative to the minimum requirement for country banks and a response of about one-half that for reserve city banks[10] The increase in note issue by country banks over the period was greater then because the decline in local interest rates was larger there, and in addition the responsiveness to changes in the local loan rate, the estimated coefficient of R, is much larger for country banks This may be due to the possibility that in rural areas the preference for notes relative to deposits was stronger and, as local interest rates declined so that it became profitable to issue notes, country banks were able to keep a larger percentage of their notes outstanding than reserve city banks, hence country banks were much more responsive in their note-issue policy than city banks In equations (2) and (3) the observation of the Indianapolis Monetary Commission that the high seasonal variability of demand in primarily rural areas discouraged note issue is also supported In the country bank regression the seasonality coefficient is large in magnitude and significantly negative, on the other hand, in the reserve city regression it is relatively small and insignificant

III. Summary

One of the unsolved problems in nineteenth-century monetary history has been why so few national bank notes were issued even though most of the time their issuance seemed to be very profitable The Goodhart explanation for this phenomenon is unsatisfactory The higher cost of note issue for city banks is very small indeed, and the increased risk is really relevant only after 1907 It does not offer a general explanation for why note issue was so low in, say, the 1890s Just as Goodhart attempted to resurrect the explanation of Spurgeon Bell, this paper represents a modified restoration of Charles Dunbar and the Indianapolis Monetary Commission

The principal reason behind the low issue of national bank notes lay in differences in the relative profitability of note issue across areas as evidenced in the wide geographical variation in note issue Southern and western banks issued hardly any notes at all in excess of the minimum requirements, in spite of the strong demand for notes in those regions, because of the high opportunity cost of note issue as measured by the local interest rate If the country bank chose to issue notes it would lose the high interest on the difference between the market price of the collateral bond and 90 percent of par value. Similarly, the general decline in interest rates over the period in rural areas made note issue a relatively more attractive prospect for banks, and as a result the percentage of maximum note issue actually in circulation increased over time

---

[10] There is the question, of course, whether differences in local profitability of note issue are the important effect or whether differences in local interest rates only pick up systematic regional differences such as population density, differences in preferences for holding notes, or average city size, all of which might influence the relative difficulty of keeping notes outstanding and thus may be the true determinants of geographical variation in note issue A pooled regression of country and reserve city banks of the same form was estimated including both regional dummies and one for reserve cities. Once more, the R coefficient was highly significant while the dummies were generally small and insignificant, contributing little explanatory power. Variations in profitability appear to be the dominant influence

References

Balestra, Pietro, and Nerlove, Marc "Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model The Demand for Natural Gas" *Econometrica* 34 (July 1966) 585–612

Bell, Spurgeon "Profit on National Bank Notes" *A E R* 2 (March 1912) 38–60

Cagan, Phillip *Determinants and Effects of Changes in the Stock of Money, 1875–1960* New York Columbia Univ Press, 1965

Davis, Lance "The Investment Market, 1870–1914 The Evolution of a National Market" *J Econ Hist* 25 (September 1965) 355–99

Dunbar, Charles *Economic Essays* New York Macmillan, 1904

Friedman, Milton, and Schwartz, Anna *A Monetary History of the United States* Princeton, N J Princeton Univ Press, 1963

Goodhart, C A E "Profit on National Bank Notes, 1900–1913" *J P E* 73, no 5 (October 1965) 516–22

James, John A "The Evolution of the National Money Market, 1888–1911" Ph D dissertation, Massachusetts Inst Tech , 1974

Lichstein, Michael "An Explanation of the Low Issue of National Bank Notes, 1875 1900" Unpublished paper, 1968

Report of the Monetary Commission of the Indianapolis Convention Chicago Univ Chicago Press, 1898

Taylor, F M "The Final Report of the Indianapolis Monetary Commission" *J P E* 6, no 3 (June 1898) 293 322

Warren, George F , and Pearson, Frank A *Prices* New York Wiley, 1933

Autocorrelated Growth Rates and the Pareto Law: A Further Analysis

Daniel R Vining, Jr.

University of Pennsylvania

Introduction

While the Pareto law[1] is a useful first approximation to many size distributions in economics, one frequently encounters in plots of these distributions on double log paper a curve that is concave downward in shape instead of the straight line predicted by the law[2] One possible source of this curvature is a violation of the first of the assumptions that Simon (1955) uses in his model to generate the Pareto law—namely, the assumption that percentage rates of growth are independent of size, otherwise known as Gibrat's law of proportionate effect If, in fact, growth rates are negatively correlated with size, then there will be more cities or firms in the intermediate size classes than would be anticipated from the Pareto law, and an abundance of middle-sized firms or cities is precisely the meaning of the curvature. Supporting evidence for this hypothesis is the fact that for populations whose size distributions have this curvature, the empirical literature does tend to report an accompanying inverse relationship between growth rates and size.[3]

I am grateful to Yuji Ijiri and Herbert Simon for valuable comments on this and a previous draft of this paper and to an anonymous referee for extensive comments and suggestions, many of which have been incorporated into the draft here published

[1] The Pareto law states that "the logarithm of the precentage of units with a size exceeding some value is a negatively sloped linear function of the logarithm of that value" (Klein 1962, p 150) Throughout this paper, I shall use the word distribution to mean the greater-than cumulative, as is customary in the literature The rank-size relationship or distribution is another name for the greater-than cumulative and will be used here interchangeably

[2] Instances of this curvature abound in the empirical literature on firm, income, and city size distributions Among the earliest to note it were Shirras (1935, p 669) and Champernowne (1953, p 347) for the income distributions of India and England For similar evidence on city and firm size distributions, see, among others, Steindl (1965, p 194) and Parr and Suzuki (1973, p 346) My dissertation contains a more extensive list of such references

[3] Negative correlations between growth rates and sizes of metropolises have been reported by a number of researchers in recent years See, for example, Berry (1973, p 104) and Leven (1973, p. 355) For similar evidence concerning firms, see Steindl (1965, pp 214-15) and Jacquemin and Cardon de Lichtbuer (1973, pp 402-3); for personal incomes, see Thatcher (1971) My dissertation contains a more extensive list of uch references.

Journal of Political Economy, 1976, vol 84, no 2]
© 1976 by The University of Chicago All rights reserved

One interpretation of this latter phenomenon in the case of firms, and the natural one to economists, perhaps, is that the increasing size of a firm acts in some manner to inhibit its rate of growth—that there is something inherent in the very nature of size that causes a progressive decline in this rate as it expands its activities The conventional description of this tendency among the firms of an economy is that they are subject to decreasing returns to scale By this interpretation, therefore, the presence of curvature in plots of firm size distributions on double log paper for an economy is nothing more than a sign of diseconomies of size in that economy

In a recent paper, however, Ijiri and Simon (1974) indirectly argue against this interpretation by demonstrating that curvature may arise in the absence of any explicit violation of Gibrat's law, an assumption that is frequently taken in the economics literature to be equivalent to that of constant rather than decreasing returns to scale According to Ijiri and Simon, curvature may result instead simply from the existence of a strong autocorrelation in growth rates over time, a fact about the growth of firms which was not allowed for in the growth process Simon (1955) originally postulated to account for the Pareto law [4] In using this fact about the growth of firms to explain the observed curvature, however (or so Ijiri and Simon seem to argue), no recourse is needed to an additional assumption of a negative correlation between growth rates and size

What I shall show here is that it is not autocorrelated growth per se but rather other features of the model in which this autocorrelation is embedded that generate the curvature Specifically, in their model autocorrelated growth is modeled in such a way that growth dies out with the increasing age of a firm (Ijiri and Simon 1964, pp 86–87, 1974, p 322) [5] If autocorrelated growth rates are not accompanied by this inverse relationship between growth rates and age but rather are modeled so that "leads and lags dissolve with time," to use the language of Nelson and Winter (1973, p. 448), then curvature of the type observed will not arise [6] This is demonstrated via the continuous model of autocorrelated growth presented in Ijiri and Simon (1967), which is really nothing more than Gibrat's model (1931) with nonindependent growth ratios I conclude that autocorrelated growth by itself cannot be the source of the observed curvature, and that therefore the latter may very well be a reliable sign of diseconomies of size in an economy (or of some characteristic closely associated with size, such as age)

---

[4] The data reveal a strong element of persistence in most economic processes. For evidence on metropolitan growth rates, see Bogue (1953, p 35) and on growth rates of firms, Ijiri and Simon (1967, p 354) and Singh and Whittington (1975, pp 21–22) Persistency in the growth rates of urban areas appears to be significantly higher than that in firm growth rates (see n. 10 below) Personal income and wealth data have been insufficiently detailed to date to determine if their year-to-year rates of change are likewise highly correlated In his revision of his master's thesis, Champernowne (1973, p 204) conjectures that the correlation between logged increments of income this year and last, for a given individual, will lie somewhere between 2 and .8 An indirect estimate of the correlation coefficient by Creedy (1974, p 411) suggests, however, that it is very small.

[5] And, therefore, since age is positively correlated with size, there will be a negative correlation between size and growth rate for any cross-sectional sample of firms. Though ceteris paribus growth rates are independent of size—"the ceteris paribus assumption being that the previous growth of the firms being compared took place at about the same time" (Ijiri and Simon 1964, p 81)—the likelihood of ceteris paribus ever holding for two different-sized firms declines with the absolute difference in their sizes

A Model of Autocorrelated Growth in which Firm Sizes Are Continuous Random Variables

In the literature on the size distribution of firms, there have evolved essentially two representations of the growth processes underlying these distributions The first follows upon a model of Simon (1955) originally developed to account for certain statistical regularities observed in the distribution of word frequencies It is this model that Ijiri and Simon have built upon in their subsequent papers on the subject of firm size distributions (1964, 1971, 1974) (An exception is their 1967 paper, which really belongs to an entirely different tradition of model, to be described below) The more elaborate models presented in Steindl (1965) may also be fairly viewed as having the basic structure of this model at their heart

It is a model in which (*a*) the sizes of firms are integer valued, and (*b*) the smallest size is normalized to one The most straightforward revision of this model to allow for autocorrelation in the growth process is that presented in Ijiri and Simon (1964, 1974) Unfortunately, the model so revised has the added feature that growth rates decline with age of firm It is therefore quite conceivable that it is this latter feature rather than autocorrelation per se which leads to curvature in the firm size distribution To determine whether this is so, it is necessary to model autocorrelation in such a way that this negative correlation between age of firm and growth rate is absent I have found no very appealing way of doing this within the discrete structure of the Ijiri-Simon model

Fortunately, there is a different tradition of model which treats the size variable as a continuous one and is probably best exemplified in the paper by Hart and Prais (1956), which followed the work of Gibrat (1931) Unfortunately, the Gibrat model is in several respects distinctly more limited than that of Ijiri and Simon

First, as Champernowne (1956) and Kendall (1956) note in their discussions of the Hart-Prais paper, the Gibrat model does not allow for firm entry at the smallest size (in contrast to that of Ijiri and Simon), possibly because it is much more cumbersome to have this feature in a continuous model. Second, it does not allow for autocorrelation in the growth process, the latest paper by Ijiri and Simon (1974) demonstrates that their model, again by contrast, is able to do so with only a minor loss of analytic tractability. One is therefore unable, by a straightforward application of the Gibrat model, to determine the extent to which the discrete model of Ijiri and Simon misrepresents the effects of autocorrelation.

---

[6] This is in conflict with the results reported by Nelson and Winter, however. See Nelson and Winter (1973, p 447)

To do so would seem to require a synthesis of the two models, which is what I shall attempt in this section. Since it is clearly desirable to treat the sizes of firms as continuous variables, in this respect I shall follow the model of Hart and Prais. On the other hand, since entries of new firms do take place at a fairly regular rate and since growth rates are serially correlated over time, I will allow firms to enter the existent population at a minimum size and then proceed to grow in accordance with a series of nonindependent draws from a growth-rate distribution, as in the Ijiri-Simon model. In contrast to the Ijiri-Simon model, however, the draws are made blind to age as well as size, so that the expected growth rate of any given firm over its future is independent of its current state. At the same time, this does not preclude the possibility, which seems to be implicit in Ijiri and Simon's exposition, that momentarily slow-growing firms will congregate in certain size classes and thereby give the impression that these sizes are inimical to growth.

Models of this form have so far resisted my attempts at analytical solution, and I have therefore sought the properties of the size distribution that they generate via simulation. The remainder of this section describes one specific model incorporating the features described above and the next the results of a simulation experiment performed on it.

The basic structure of the autocorrelation component of this model, interestingly enough, has already been worked out by Ijiri and Simon in their 1967 paper, although the approach was not pushed toward an explicit treatment of the steady-state distribution of firms implied by it.[7] In what follows I have borrowed heavily from that paper, in some places using their own language. I have elected not to disentangle what is theirs from what is mine, in the interest of readability.

Let S_{it} be the size of the ith firm at time t and r_{it} the growth ratio that relates S_{it} to $S_{i(t-1)}$, so that

$$S_{it} = r_{it}S_{i(t-1)} \qquad (1)$$

Let us further decompose r_{it} into two factors: one is a growth factor applicable to the ith firm only (the individual growth factor), ρ_{it}, and the other a growth factor that affects equally all firms in the economy, $\bar{\rho}$. Thus,

---

[7] In Ijiri and Simon (1974, p. 322), the empirical results of the 1967 paper are presented as confirming the assumptions of the model presented in the 1964 and 1974 papers. For those familiar with these papers, the purpose of the present paper may be thought of as an exegetic one in which the two models advanced in these papers are shown to be of two quite different growth processes.

$$r_{it} = \rho_{it}\bar{\rho} \tag{2}$$

Hence,

$$S_{it} = \rho_{it}\bar{\rho}S_{i(t-1)} \tag{3}$$

Assume that the individual growth ratio, ρ_{it}, in the tth period is the product of the growth ratio of the same firm in the $(t - 1)$ period and a random factor, ε_{it}, which is distributed independently and identically for every firm and for every t—that is,

$$\rho_{it} = \varepsilon_{it}\rho^{\beta}_{i(t-1)}, \tag{4}$$

where β is some constant in the range $0 \leq \beta < 1$, and ε_{it} has mean 1 and variance σ^2

If $\beta = 0$, then we have the familiar Gibrat model that Hart and Prais have employed in their work On the other hand, even if $\beta > 0$ (growth rates are positively correlated over time), (4) together with (3) is still consistent with Gibrat's law since the expected growth ratio over the future is independent of current firm size and age That is, since $\beta < 1$, growth rates will have a tendency to regress toward the average growth rate of the entire economy, $\bar{\rho}$

To complete the description of the model, there remains only the question of entry, ignored in the Hart and Prais model as well as in the model of Ijiri and Simon (1967) just presented In the latter's discrete model, however, firms enter at the minimum size, 1, at some constant rate, α This rate is measured empirically (see Simon and Bonini 1958, p 611, Engwall 1968, pp 142-43, Vining 1974, p 327) as the share of total growth for the firms above a given minimum size over a given time period which is due to new firms entering this population during this time period By the notation introduced above, the total growth due to old firms, OG_t, between the beginning and the end of time t is given by

$$OG_t = \sum_{i=1}^{k(t-1)} [S_{it} - S_{i(t-1)}], \tag{5}$$

where $k(t - 1)$ is the number of firms in the economy at the end of time $(t - 1)$ If the share of total growth by new firms is α, then the total number of new assets to be added to the economy via new firms should be, on the average,

$$NG_t = OG_t[\alpha/(1 - \alpha)] \tag{6}$$

New firms are entered at the end of time t The number of such firms is given by

$$NC_t = NG_t/S_{\min}; \tag{7}$$

where $S_{\min}$ is the minimum size. Since NC_t must be an integer, in the actual simulation of this process, it is rounded off to the nearest integer and the remainder (negative if rounded up and positive if rounded down) is carried over to the next time period and added to NC_{t+1}. Thus, in the actual simulation, NC_t is calculated exactly by

$$NC_t = INT[(NG_t/S_{\min}) + .5 + (R_{t-1}/S_{\min})], \qquad (8)$$

where, $R_{t-1} = NG_{t-1} - (NC_{t-1}S_{\min})$, R_0 equals zero, and INT is the Fortran function that converts the number in its argument consisting of a whole number plus a fraction to that whole number. For the next time period $t + 1$, these new firms are regarded as old firms with size $S_{\min}$ and growth ratios $\rho_{i(t+1)}$, calculated from (4) with ρ_{it} equal to one.

The simulation proceeds, therefore, in two steps. At the first, each existent firm's growth ratio, ρ_{it}, is computed via (4), β is held constant throughout the simulation, and ε_{it} is drawn from a lognormal distribution with mean 1 and variance σ^2.[8] Each firm is then "grown" to its next period's size and the total increment of assets for the old firms, OG_t, is computed via (5). At the second step, the number of new firms to be added is computed via (6) and (8), their initial growth ratios are assumed to be one in the calculation of their growth rates in the succeeding time period, while their initial sizes are taken to be the minimum firm size, $S_{\min}$. The simulation executes these two steps for a large number of iterations. At the end, the existent firms are arrayed in order of decreasing size and plotted against their ranks on double log paper. The next section of this paper presents these plots for different values of β with σ, $\bar{\rho}$, and α held constant.

Rank-Size Relationships Generated by the Model

This section compares the rank-size relationships generated by simulation runs of the model described in the last section for different degrees of autocorrelation, that is, for different values of β. All other parameters of the model ($\bar{\rho}$, α, and σ) were held constant during these runs. Specifically, $\bar{\rho}$ was set equal to 1.0285, σ to 0.0095, and α to 0.05. The particular values of the first two were chosen simply to insure that firms rarely decline in size. It will be remembered that the Ijiri-Simon model is one of a pure growth process, and therefore in order to make a valid comparison with that model the one presented here must also possess that property. Although the value for α is somewhat lower than that reported in the literature,[9] higher values were found to demand too much computing time for the number of time periods that I wanted to run the simulation. Run time is a direct function of the number of firms in the population, which are in turn growing at the (exponential) rate of $(\bar{\rho} - 1)/(1 - \alpha)$. With $\bar{\rho} = 1.0285$ and $\alpha = 0.05$, the aggregate rate of growth is 3 percent per time period.

---

[8] This implies lognormally distributed growth rates, i.e., a distribution of growth rates skewed to the right. The empirical literature tends to substantiate this assumption. See Madden (1956, p. 250) for the evidence concerning cities and Hart and Prais (1956, pp. 170–71) for that concerning business firms. Mandelbrot (1963, pp. 30–31), however, argues that the distribution of logged growth ratios is closer to a two-tailed Pareto than the normal distribution that one would expect from lognormally distributed growth rates—i.e., that there are many more extreme observations than one would expect from a normal distribution.

It is customary in simulation studies to test the procedure on a special case of the model for which there is a known analytic solution. In the present case, such a case occurs when $\beta = 0.0$. That is, one would anticipate from the analytic results presented in Simon (1955) for a nonautocorrelated growth process that the steady-state relationship between rank and size would trace out an approximately straight line on double log paper with an absolute slope of $(1 - \alpha)$, or .95 if $\alpha = 0.05$, for $\beta = 0.0$. The first graph of figure 1 (for $\beta = 0.0$) bears out this expectation. The rank-size relationship, after 300 years of simulated time, closely conforms to the predictions of the Simon model: it is linear on double log paper with absolute slope, $1 - \alpha$, or .95. (There are divergences from linearity in the upper end, which are due to the fact that the slope converges only asymptotically to $[1 - \alpha]$ from above [see Vining 1975]. This means a somewhat larger slope than $[1 - \alpha]$ initially; therefore, as observed in fig. 1, the points will lie above the predicted rank-size relationship for large S and small R.) On the basis of this result, I conclude that the simulation procedure is an accurate embodiment of that model. However, the reader is invited to inspect the simple Fortran program used to implement the procedure, available from me on request.

To compare the rank-size relationships for different degrees of autocorrelation in the growth process, I ran the model for four different values of β: 0.3, 0.5, 0.7, and 0.9.[10] Typical log-log charts of the rank-size relationships generated by the model for each of these four values of β, again after 300 years of simulated time, are presented in the remaining graphs of figure 1. While there is evidence, in agreement with the Ijiri-Simon results (1974), that the degree of curvature increases with the degree of autocorrelation—that is, with higher β—the curvature is convex rather than concave and lies on the down side of the Pareto curve rather

---

[9] See, for firms, Simon and Bonini (1958, p. 612) and Engwall (1968, p. 153); for cities, see Vining (1974, p. 327).

[10] For two 4-year periods, Ijiri and Simon (1967, p. 354) estimate β to be around one-third for the largest 100 corporations in the United States, while I have estimated β for American urban areas, for two 10-year periods, to be around .75 (see Vining [1975] for a

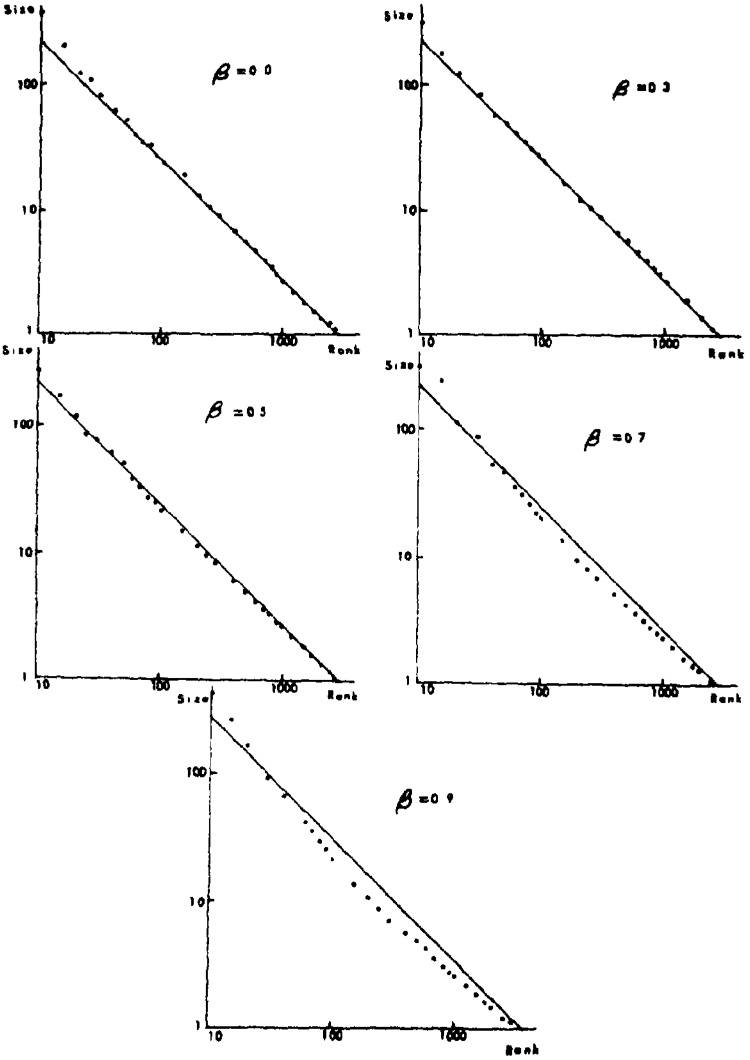


FIG 1.—Typical rank-size relationships generated by the simulation for β = 0.0, 0.3, 0.5, 0.7, and 0.9 (The solid line is the rank-size relationship predicted by the Simon model [1955] and is drawn with absolute slope .95 from the smallest size.)

than the upper as reported by Ijiri and Simon Convex curvature generally reflects an underlying positive correlation between growth rates and size the larger number of firms generated when $\beta = .9$ is also a sign of a positive correlation between growth rates and size [11]

It seems clear, therefore, that autocorrelation by itself (i e , with the expected long-run growth rate for a given firm independent of its current age and size) cannot be the source of the curvature actually observed in empirical rank-size relationships in that it cannot be the source of the negative correlation between growth rates and size which is the immediate cause of this curvature. Quite the opposite, autocorrelation appears to generate a positive correlation between size and growth rate, at least for a high degree of autocorrelation in the growth process, and therefore a higher degree of concentration than would occur if growth rates were independent over time [12] I conclude that for autocorrelated growth rates to be compatible with the observed curvature, the growth rates for a given firm must in addition have expected values that are a declining function of that firm's size or of some variable highly correlated with size, such as age Of course, to constrain growth rates in this fashion is to rob the original hypothesis (i e , that autocorrelation in the absence of

---

description of the data and fitting procedures) In order to compare the magnitudes of these two estimates of β, it is necessary to standardize them for the same time period After adjusting both estimates to an annual basis (by a method easily adopted from Telser [1967], see Vining [1975]), β turns out to be 65 and 95 for firms and urban areas, respectively Thus, if one were to regard the year as the "natural" time period for most economic processes, β for two of these processes appears to be well over one-half These estimates, of course, are intended only to give the reader some rough idea of what degree of autocorrelation there is in two important processes in economics, city and corporate growth, so that he can look at the charts in figure 1 more intelligently

[11] They are typical in the following sense for all simulations, convexity in the rank-size charts appears with $\beta = 5$ and increases thereafter The simulations for which the charts represent a typical outcome were replicated three times (i e , with three different random number seeds) for three different initial size distributions (one of 10 firms with sizes arrayed in rank-size order with slope parameter equal to 95, one of three firms, again in rank-size order, and one of two firms, of sizes five and three, respectively)

[12] It is interesting to note that Steindl (1965, pp 213–14) has argued on intuitive grounds that persistency in the growth process will generate this positive correlation and therefore a higher degree of concentration than would be normal if growth rates were independently distributed over time Another interpretation of the increased convexity with increasing β has been offered by the referee "As $\beta \to 1$, the AR(1) process, (4), approaches nonstationarity In fact, it approaches a random walk Thus, for large β, we would expect the behavior of the series to be quite different from that for β near 0 In particular, convergence to a steady-state distribution may be considerably slower It is quite conceivable that the increased convexity with increasing β is due to this. You have already noted that the slight convexity at the upper end for $\beta = 0$ is due to the convergence to the asymptotic relationship from above I can find nothing in your figure which would be inconsistent with the hypothesis that the observed convexity represents a shifting of the effect already noted for $\beta = 0$ Convergence to a steady-state distribution can be extremely slow in such processes. If true, the slow convergence could be an essential and important aspect of the model "

any such constraint will generate a negative correlation coefficient computed on a cross-sectional sample of growth rates and firm sizes) of much of its intrinsic interest Naturally, if growth rates are ever declining, as they are if they are negatively correlated with age, they must also be strongly autocorrelated However, it is the former and not the latter feature of the model that is the source of the observed curvature [13]

A striking implication of this finding is that the rank-size rule, or Pareto's law, must then exist, where it does indeed exist, in a kind of balance between autocorrelated growth rates and the inhibitory effects of age on growth rates the latter pull the distribution into concavity, the former into convexity To judge from the empirical city and firm size distributions, however, the age effect appears to have recently come to dominate the autocorrelation effect

Conclusion and Summary

This paper had the single objective of demonstrating that autocorrelation in firm growth rates cannot by itself account for the negative correlation between growth rates and size frequently observed in cross-sectional samples of corporations (as well as cities) and therefore cannot in turn account for the concave curvature observed in the distribution of their sizes on double log paper which is an expression of this negative correlation The method used to demonstrate this fact was to simulate the continuous model of Gibrat with the added features of autocorrelated growth rates and expanding firm population The resultant distribution of firm sizes showed convexity rather than concavity, implying a positive correlation between firm size and growth rates Thus, it appears that the negative correlation between firm size and growth rates must be built directly and explicitly into the model if curvature like that observed is to be generated In short, it cannot arise as a by-product of some weaker set of assumptions concerning firm growth

References

Berry, B J. L "Contemporary Urbanization Processes " In *Geographical Prospectives and Urban Problems* (a symposium organized by the Committee on Geography of the Division of Earth Sciences, National Academy of Sciences, 1971), edited by Frank E Horton. Washington Nat Acad Sci , 1973

---

[13] Table 1 of Bogue (1953, p 13) clearly shows this effect of age on standard metropolitan area (SMA) growth rates Cities entering the SMA population at each decade (i e , surpassing 100,000 population) show a higher average growth rate than those which entered in a previous decade and a lower average growth rate than those which entered in a later decade Williamson and Swanson (1966, p 53), in a multiple-regression study of the effects of size and age on city growth rates in the nineteenth century in this country, report a negative effect of age and no effect of size That is, for a given size, the older the city, the lower the growth rate, on the average, while, for a given age, size has no effect on a city's growth rate Bock and Farkas (1973, pp 24–34) present similar evidence for American corporations

Bock, B., and Farkas, J. *Relative Growth of the Largest Manufacturing Corporations, 1947–1971, Subsets from an Unknown Set.* New York: Conference Board, 1973.

Bogue, D. J. *Population Growth in Standard Metropolitan Areas, 1900–1950* Washington: House and Home Finance Agency, Office of the Administrator, Division of Housing Res., Government Printing Office, 1953.

Champernowne, D. G. "A Model of Income Distribution." *Econ. J* 63 (June 1953). 318–51

———. "Discussion on Paper by Mr Hart and Dr Prais" *J. Royal Statis. Soc.* A119 (1956) 181–82.

———. *The Distribution of Incomes between Persons* London Cambridge Univ Press, 1973

Creedy, J, Jr "Income Changes over the Life Cycle" *Oxford Econ. Papers* 26 (November 1974) 405–23

Engwall, L. "Size Distribution of Firms—A Stochastic Model" *Swedish J Econ.* 70 (September 1968) 138–59

Gibrat, R *Les Inégalités économiques* Paris Librairie du Recueil Sirey, 1931

Hart, P E, and Prais, S J "An Analysis of Business Concentration" *J Royal Statis Soc* A119 (1956) 150–81

Ijiri, Y, and Simon, H A "Business Firm Growth and Size" *A E R* 54 (March 1964) 77–89

——— "A Model of Business Firm Growth" *Econometrica* 35 (April 1967) 348–55

——— "Effects of Mergers and Acquisitions on Business Firm Concentration." *J P E* 79, no 2 (March/April 1971) 314 22

——— "Interpretations of Departures from the Pareto Curve Firm-Size Distributions" *J.P.E* 82, no 2 (March/April 1974) 315–31

Jacquemin, A P, and Cardon de Lichtbuer, M "Size Structure, Stability, and Performance of the Largest British and EEC Firms" *European Econ Rev* 4 (December 1973) 393 408

Kendall, M G "Discussion on Paper by Mr Hart and Dr Prais" *J Royal Statis Soc* A119 (1956) 184 85

Klein, L *An Introduction to Econometrics* Englewood Cliffs, N J Prentice-Hall, 1962

Leven, W I "A Markov Approach to the Optimum Size of Cities in England and Wales" *Urban Studies* 10 (October 1973) 353-66

Madden, K 'Some Indications of Stability in the Growth of Cities in the United States" *Econ Development and Cultural Change* 4 (1956) 236 52

Mandelbrot, B "A Survey of Growth and Diffusion Models of the Law of Pareto" I B M Research Report NC-253, May 1963

Nelson, R R, and Winter, S G "Toward an Evolutionary Theory of Economic Capabilities" *Papers and Proc American Econ Assoc* 63 (May 1973) 440–49

Pari, J, and Suzuki, K "Settlement Populations and the Lognormal Distribution" *Urban Studies* 10 (October 1973) 335 52

Shirras, G F "The Pareto Law and the Distribution of Incomes" *Econ J* 45 (December 1935) 663 81

Simon, H A "On a Class of Skew Distribution Functions" *Biometrika* 42 (December 1955) 425 40

Simon, H A, and Bonini, C P "The Size Distribution of Firms" *A E R* 48 (September 1958) 607 17

Singh, A, and Whittington, G "The Size and Growth of Firms." *Rev Econ Studies* 42 (January 1975) 15 26

Steindl, J *Random Processes and the Growth of Firms* New York Hafner, 1965

Telser, L G "Discrete Samples and Moving Sums in Stationary Stochastic Processes" *J American Statis Assoc.* 62 (June 1967) 484 99

Thatcher, A. R. "Year-to-Year Variations in the Earnings of Individuals." *J. Royal Statis. Society* A134 (1971) 133–38.

Vining, D. "On the Sources of Instability in the Rank-Size Rule: Some Simple Tests of Gibrat's Law " *Geographical Analysis* 6 (October 1974): 313–29.

——— "Models of Urban and Spatial Concentration." Ph D. dissertation, Carnegie-Mellon University, 1975.

Williamson, J G , and Swanson, J. A. "The Growth of Cities in the American Northwest, 1820–1870." *Explorations in Entrepreneurial History 2nd Series* 4 (suppl. 1966) 1–101.

It Takes Two to Tango, or *Sind* "Separable Externalities" *Überhaupt Möglich?*

William J. Baumol

Princeton and New York Universities

Various things that would *seem* to require a pair of participants can apparently be carried out alone. We have victimless crimes and artificial insemination. Yet it seems hard to think of an externality without (at least) two participating parties: a victim (beneficiary) and a generator (even a victimless crime requires a perpetrator). I will argue that under these circumstances the phenomenon called a "separable externality" by Davis and Whinston is simply impossible.

It will be recalled that Davis and Whinston define an externality to be separable if the function describing the cost (benefit) of the externality to the victim (beneficiary) can be written as the sum of two terms, one not depending on the behavior of the generator of the externality and the other independent of the behavior of his victims. This independence condition was suggested by these authors as a case in which it may be feasible to calculate the marginal damage figure on which a Pigouvian tax must be based. For where externalities are inseparable, the magnitude of that marginal damage figure will depend on the activity levels of the victims, which, incidentally, are clearly beyond the control of the generator.

Since the appearance of the Davis-Whinston contribution the concept of separable externalities has recurred at least sporadically in the literature where it seems implicitly to be suggested that it is a very real phenomenon. I will argue here that it is not merely unlikely but, strictly speaking, impossible.

I. The Postulates

Let x_1 and x_2, respectively, represent the activity level of the victim (individual 1) and of the generator (individual 2) of the externality. Then

Work on this paper was supported by the National Science Foundation through the project, the Economics of the Environment, at Princeton and New York Universities. I am also grateful to Herbert Mohring and George Stigler for their helpful expository comments.

[*Journal of Political Economy*, 1976, vol. 84, no. 2]

© 1976 by The University of Chicago. All rights reserved.

we take the function $E(x_1, x_2)$ to be a measure of the magnitude of the external effect, presumable as judged from the viewpoint of the victim.

Now, separability may be defined by the condition that

$$E(x_1, x_2) = E^1(x_1) + E^2(x_2), \tag{1a}$$

or, writing $E_{12} \equiv \partial^2 E/\partial x_1\, \partial x_2$, etc.,

$$E_{12} = E_{21} = 0, \tag{1b}$$

if these derivatives exist.

Next we write the following relationships to express the assumption that it takes two to make an externality:

$$E(x_1, 0) = E(0, x_2) = 0 \text{ (all } x_1, x_2) \tag{2}$$

$$E(x_1^*, x_2^*) \neq 0 \tag{3}$$

for some nonzero values x_1^*, x_2^*.

Condition (3) is obviously no more than the assertion that the externality exists. The meaning of our fundamental condition (2), which concerns the magnitude of the *activity* levels of the two groups, can be made clear by two examples. The first is the standard illustration of the electricity generator whose smoke raises the costs of a nearby laundry. Obviously, there will be no such damage if either the electric company ceases operation ($x_2 = 0$) so that smoke is no longer emitted or the laundry no longer washes anything ($x_1 = 0$) so that there are no laundry costs to be increased. But though condition (2) depends on the activity levels of both parties it does not necessarily require the generators (or the victims) of the externality to "do" anything. Their mere presence may be sufficient to give rise to the problem, and the magnitude of the pertinent "activity" is then indicated by the number of persons present. Thus, consider a room occupied simultaneously by a number (x_2) of malodorous persons and a number (x_1) of persons offended by the odor. Then the externality clearly disappears if $x_1 = 0$ because no one is there to suffer the displeasure. It disappears also if $x_2 = 0$ so that no body odor is given off. Hence $E(0, x_2) = E(x_1, 0) = 0$, as asserted.

II. The Formal Argument

I now undertake to prove:

If an externality is separable (satisfying condition [1a] or [1b]) it is impossible for it also to satisfy conditions (2), the requirement that it have both a generator and a victim, and condition (3), existence of the externality.

Proof:
From (2), for $x_1 = 0$ we have

$$\frac{\Delta E(0, x_2)}{\Delta x_2} \equiv \frac{E(0, x_2^*) - E(0, 0)}{x_2^* - 0} = 0, \tag{4}$$

while by (3), for $x_1 = x_1^*$

$$\frac{\Delta E(x_1^*, x_2)}{\Delta x_2} \equiv \frac{E(x_1^*, x_2^*) - E(x_1^*, 0)}{x_2^* - 0} \neq 0 \tag{5}$$

Hence, subtracting (4) from (5)

$$\frac{\Delta(\Delta E/\Delta x_2)}{\Delta x_1} = \frac{\Delta E(x_1^*, x_2)/\Delta x_2 - \Delta E(0, x_2)/\Delta x_2}{x_1^* - 0} \neq 0.$$

If $\partial^2 E/\partial x_2\, \partial x_1$ exists throughout, it follows by the mean-value theorem that there is some value x_2^{**}, with $0 \leq x_2^{**} \leq x_2^*$, at which $\partial E/\partial x_2\, \partial x_1 \neq 0$, thus contradicting (1b) *Q E D*

But even where that derivative E_{21} does not exist it is easy to show that (1a) contradicts (2) and (3) For, by (3)

$$E(x_1^*, x_2^*) = E^1(x_1^*) + E^2(x_2^*) \neq 0 \tag{6}$$

Hence, at least one of the terms must be nonzero Without loss of generality, take it to be the first term Then, assuming $E^1(0) = E^2(0) = 0$,

$$E(x_1^*, 0) = E^1(x_1^*) + E^2(0) \neq 0, \tag{7}$$

clearly contradicting (2)

In other words, a strictly separable externality must be a queer case in which damage is produced by a nonexistent perpetrator or is sustained by a nonexistent victim[1]

III. Separability Almost Everywhere: Concept and Examples

As a substitute for the concept of strict separability, as defined by (1), it may be useful to consider a weaker case, which I will call "local separability," the case where the cross partial is zero for some set of values of the variables

$$\partial^2 E/\partial x_1\, \partial x_2 = 0 \ (x_1 \in S_1,\ x_2 \in S_2) \tag{8}$$

It is not easy to think of examples of externalities that are even locally separable All the classic externality illustrations, such as the effect of smoky electricity production on a nearby laundry, are clearly inseparable throughout That is, the larger the volume of emissions the greater will be the *marginal* cost of laundry One can, however, construct some cases in which the external damage is, in effect, a fixed cost, affecting the victim's marginal costs only where his activity level approaches zero. This can occur when the damaging activity causes the victim to adopt some shielding device and the size of the investment in shielding is unaffected by the scale of the victim's activities. In such circumstances the external damage is independent of the scale of the victim's activities but dependent on the victim's decision on exit or entry into the affected activity. The externality is separable not only for a small neighborhood, but *almost everywhere*, that is, everywhere but on one axis.

Thus, imagine a writer who has bought an isolated property and a neighbor builds a house next door, which makes the writer too nervous to write. He decides to build a fence around his property. In this case the magnitude of the cost is the same whether the writer produces 2, 10, or 50 pages per day. It is presumably also independent of the number of neighbors who move in next to him, but that is not necessary for the separability property. In this case the writer's (external and internal) cost function takes the form

$$E(x_1, x_2) = E^1(x_1) + \delta k \qquad \delta = \begin{cases} 0 \text{ if } x_1 = 0 \text{ or } x_2 = 0 \\ 1 \text{ if } x_1 > 0 \text{ and } x_2 > 0 \end{cases} \tag{9}$$

where x_1 = the number of pages written, x_2 = the number of neighbors and k is the cost of erecting the fence.

An illustration somewhat more conventional in appearance arises where the neighbors are noisy, and the amount of the noise (and hence the cost of the required sound insulation) varies with x_2, the number of neighbors, but not with the volume of the writer's production. In that case the relationship becomes

$$E(x_1, x_2) = E^1(x_1) + \delta E^2(x_2) \tag{10}$$

where δ is defined as before.

But notice that because of the definition of the δ (i.e., its dependence on the presence of some victim-activity) even this case is one of local separability.[1]

IV. Strict Separability within the Feasible Region

For reasons which will become clear in the concluding section, it may be useful also to distinguish cases in which nonseparability occurs in principle but is irrelevant because the region where separability is violated is simply not feasible. Another artificial illustration will bring out the point.

---

[1] One may try to make an exception of the case in which the writer gives up his activity but is driven to put in sound insulation in any event. Here x_1 must be redefined since it is not the writing but the writer who is now adversely affected by the neighbor's noise. If we define x_1 as the amount of time he spends on his property it becomes clear that this is still a case of local separability.

Consider some pollutant that affects the atmosphere equally throughout the world, in terms of its effects on people who absorb it. This might occur if these effects take place whenever the quantity absorbed exceeds a certain level and then are not aggravated by any additional absorption. Suppose, moreover, that an individual is able to protect himself from the damage by some sort of measure involving a fixed expenditure (air conditioning, by daily ingestion of an antidote pill, or by wearing a gas mask) The externality is then, obviously, separable locally ($E_{12} = 0$ except in the vicinity of $x_1 = 0$, or $x_2 = 0$) Moreover, since the cost can be avoided altogether only by suicide, the nonseparable portion of the function can be taken to be infeasible

Zero operations may also not be feasible for a firm, notably for a public utility which is required by law to supply a given service even though it would prefer to exit from that line of activity (message telegraph and passenger railroads are well-known examples) If such a firm is either the victim *or the generator* of a locally separable externality, that externality, too, may be strictly separable thoughout the feasible region

In sum, it is indeed possible to think of all sorts of cases that approximate separability to one extent or another, but plausible and significant illustrations do not seem easy to come by

V. Relevance of the Separability Issue

As was already noted, the concept of separability was first applied to externalities by Davis and Whinston in order to characterize the circumstances in which marginal damage costs are, at least in principle, comparatively easy to determine They pointed out that where externalities are inseparable, the magnitude of the emitter's *marginal* damage is beyond his control since it depends on the behavior of his victims Consequently it is difficult to decide on the magnitude of a tax on emissions if that tax, on good Pigouvian principles, is to be set equal to marginal damage But where externalities are separable this difficulty can be avoided [2]

This exception obviously holds equally for externalities that are separable locally, especially where they are separable almost everywhere, as they are in the illustrations of Section III Thus, the implications of this note for the Davis-Whinston analysis do not seem very substantial.

However, the discussion is rather more pertinent for some other uses of the concept Recently, for example, John R. McNamara in an unpublished note has used it in a discussion of the convexity implications of externalities, arguing that the tendency of externalities to introduce nonconvexities into the production set applies only in the nonseparable case.[3] The connection between detrimental externalities and nonconvexity is easy to illustrate in the case of a linear technology. Suppose the production set is composed of the points that satisfy

---

[2] Dusansky and Kalman (1972) have recently argued that the class of cases which escape the difficulty is considerably broader than Davis and Whinston (1962) suggested

$$x_2 \leq a - bx_1 + E(x_1, x_2), \qquad x_1 \geq 0, x_2 \geq 0 \tag{11}$$

Then, if there were no externality so that $E(x_1, x_2) \equiv 0$, the production frontier would be linear and the production set obviously convex. Next, suppose instead that the externality is present and detrimental, so that for any feasible x_1, x_2 $E(x_1, x_2) \leq 0$ Let x_1^*, x_2^* be some point *on the frontier* at which the externality is nonzero so that

$$E(x_1^*, x_2^*) < 0 \tag{12}$$

Let L denote the line segment given by $x_2 = a - bx_1, x_1 \geq 0, x_2 \geq 0$ By (2) and (11) the points $(0, a)$ and $(a/b, 0)$ (i.e , the two intercepts of line L) lie in the possibility set as well as in L and by (2) are also the fixed intercepts of the possibility set (11) The nonconvexity then follows from the fact that there must be some point on L between these two points which does not lie in the possibility set. For since (x_1^*, x_2^*) lies in the frontier of that set, for any point (x_1^*, x_2) in that set we must have $x_2 \leq x_2^*$ Therefore there is no value of x_2 for which (x_1^*, x_2) lies in the line L, since by (11) and (12)

$$x_2 \leq x_2^* \leq a - bx_1^* + E(x_1^*, x_2^*) < a - bx_1^*,$$

which proves the nonconvexity of the production set

Where an externality is strictly separable this argument breaks down because the endpoints are no longer unaffected by the presence of the externality That is, if $\hat{x}_1$ and $\hat{x}_2$ are the intercepts of that frontier we have by (1a), $\hat{x}_2 = a - b0 + E^1(0) + E^2(\hat{x}_2) = a + E^2(\hat{x}_2) \neq a$

In diagrammatic terms (fig 1), where it is completely separable the presence of the externality simply reduces the production set from *OAB* to *OAB′*, that is, from one convex region to another However, if an externality is merely separable locally, even if it is separable almost everywhere, as in (9) and (10), the nonconvexity is unavoidable Since in this case the externality vanishes (only) on the x_2 axis, point *B* rather than *B′* is the intercept of the possibility frontier even when the introduction of such an externality affects every other point in the production set. That is, the production set then becomes the nonconvex region *OAB′B* whose northeast boundary is the heavy line *BB′A* This may bring with it all of the problems of pricing and multiple maxima which can arise in the presence of nonconvexity

---

[3] For further discussion of this connection between externalities and nonconvexities, see Baumol and Oates (1975), chap 8

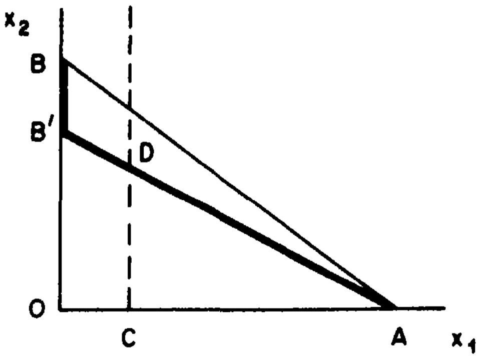


FIG 1

However, where the function is strictly separable *in the feasible region*, that is, where the nonseparable portions of the set are also not feasible, it is true that the convexity problem can disappear For example, if in figure 1 any value of activity level x_1 less than OC is not feasible, the production set with the externality present becomes the convex region CAD How important this exception is likely to be in practice, the reader will have to judge for himself

References

Baumol, William J , and Oates, Wallace E *The Theory of Environmental Policy*. Englewood Cliffs, N J Prentice-Hall, 1975

Davis, Otto A , and Whinston, Andrew "Externalities, Welfare, and the Theory of Games " *J P E* 70, no 3 (June 1962) 241–62

Dusansky, R , and Kalman, P J "Externalities, Welfare and the Feasibility of Corrective Taxes " *J P E* 80, no 5 (September/October 1972) 1045–51

ote on Tariffs, Nontariff Barriers, and bor Protection in United States anufacturing Industries

hn H Cheh

reasury Board Secretariat, Canada

I. Introduction

This note tests empirically whether all United States trade restrictions, both tariff and nontariff, protect labor in manufacturing industries The relationship between United States tariffs and labor intensity as measured by either direct labor use or labor cost has been examined by Vaccara (1960) and Basevi (1966) This paper's point of departure lies in taking *nontariff barriers* explicitly into account in a "total" measure of protection

Due partly to the success of previous rounds of multilateral tariff negotiations in reducing nominal tariffs, nontariff barriers to trade have been receiving increasing attention in recent years from economists and policymakers alike They are emerging as a focal point of discussion in the current Tokyo Round of multilateral tariff negotiations As Baldwin notes in his study of nontariff barriers, "the lowering of tariffs has, in effect, been like draining a swamp The lower water level has revealed all the snags and stumps of non-tariff barriers that still have to be cleared away" (Baldwin 1970, p 2) Ohlin (1969) has developed the thesis that success with tariff reductions may generate pressures on governments to impose new forms of nontariff barriers The increasing importance of nontariff barriers for the United States in both relative and absolute terms has been noted elsewhere by the author (Cheh 1974*a*, pp 328–29)

II. The Vaccara-Travis-Basevi Controversy

Vaccara (1960) has found that *nominal tariff rates* for United States manufacturing industries vary positively with labor intensity, as measured by either direct labor use (the ratio of average number of employees to value

---

This note is condensed from my doctoral dissertation (Cheh 1974*b*) It contains work completed prior to my joining the Treasury Board Secretariat of Canada I am greatly indebted to my thesis supervisors, Jagdish Bhagwati and Charles Kindleberger, for helpful guidance and comments Thanks are also due to an anonymous referee and to the editor of this *Journal* for helpful suggestions I am solely responsible, of course, for any remaining errors

Journal of Political Economy, 1976, vol 84, no 2]
© 1976 by The University of Chicago All rights reserved

of shipments) or direct labor cost (the ratio of total wages and salaries to value of shipments) This finding was confirmed by Basevi (1966). But in examining the relationship between effective rates of protection and labor intensity, Basevi obtained the following results· (*i*) no statistically significant relationship between labor intensity and effective rates of protection on total value added, (*ii*) for 1958, a statistically significant negative relationship between labor intensity and effective rates of protection on value added by labor

Based on Vaccara's finding of a positive relationship between nominal tariff rates and labor intensity, Travis (1964) has argued that the Leontief paradox may be caused by United States trade policy According to Travis, United States nominal tariffs, by restricting labor-intensive imports, is responsible for Leontief's finding that United States exports are more labor-intensive than United States imports Under free trade, labor-intensive manufactured imports will increase relative to the import of capital-intensive raw materials such as crude petroleum, paper pulp, and mineral products, thereby establishing the pattern of trade according to United States' comparative advantage in capital-intensive industries

The chain of reasoning in Travis's explanation of the Leontief paradox is critically dependent on a positive relationship between a protective measure which restricts imports and labor intensity Basing his case on Vaccara's finding, Travis therefore assumes that high nominal tariffs restrict imports This is where Basevi disagreed Basevi (1966) argued that it is effective rates that actually matter for protection, and since no positive relationship is found to exist between effective tariff rates and labor intensity, Travis's hypothesis is refuted

III. Our Hypothesis

In fact, it must be recognized that the relevant test for Travis's hypothesis is whether a total measure of protection, including both tariffs and nontariff barriers, is positively related to labor intensity [1] Since both nominal and *effective rates of protection impinge on imports* the *former* by discouraging domestic consumption of imported goods, the latter by encouraging domestic production—there is no clear-cut answer as to which is a superior index of import restriction Both nominal and effective rates will be used in our empirical analysis

Data for rates of protection are obtained from Baldwin (1970) His calculations of protective rates due to nontariff barriers included import quotas, the American Selling Price system of customs evaluation, federal excise taxes, the federal transportation tax, state and local retail taxes, and the federal subsidy to highway transportation This is not, as Baldwin

---

[1] It should be noted that Basevi (1966) included the tariff equivalent of selected quotas in his effective rates of protection There are, however, many other forms of nontariff barriers See Baldwin (1970) for a comprehensive list of nontariff distortions

TABLE 1

REGRESSION RESULTS: RATES OF PROTECTION ON LABOR INTENSITY

	Independent Variables				
Dependent Variable	Constant	Labor Use	Labor Cost	R^2	No of Observations
1 Nominal tariffs	5.73 (2.61) (2.20)	0.13* (0.05) (2.38)		0.10	51
2 Nominal tariffs	8.47 (3.16) (2.68)		0.12 (0.12) (1.03)	0.02	51
3 Nominal tariffs and nontariff barriers	9.49 (3.10) (3.06)	0.07 (0.06) (1.03)		0.02	51
4 Nominal tariffs and nontariff barriers	14.45 (3.62) (3.99)		−0.08 (0.13) (−0.56)	0.01	51
5 Effective tariffs	8.00 (7.15) (1.12)	0.26** (0.15) (1.75)		0.06	51
6 Effective tariffs	14.41 (8.49) (1.70)		0.21 (0.32) (0.66)	0.01	51
7 Effective tariffs and nontariff barriers	10.13 (8.09) (1.25)	0.22 (0.17) (1.32)		0.03	51
8 Effective tariffs and nontariff barriers	20.70 (9.52) (2.17)		−0.02 (0.35) (−0.06)	0.0001	51

SOURCES.—Rates of protection in 1964 constructed from Baldwin (1970). Labor use = no. of employees per million dollars of value of domestic shipments in 1963; Labor cost = total wages and salaries per hundred dollars of value of domestic shipments in 1963 (constructed from U.S. Bureau of the Census 1966).

NOTE.—Estimated coefficients followed by standard errors and *t*-statistics (estimated by ordinary least squares).

* Significant at 5% level.

** Significant at 10% level.

pointed out, a complete list of all nontariff distortions—various administrative and technical regulations in customs valuation, government procurement policies, and regional development subsidies, for example—have not been included because their quantitative effects on an industry-by-industry basis cannot be estimated. In spite of this qualification, Baldwin's rates are still the most comprehensive ones available at the moment.

The results of our regressional analysis are shown in table 1. In equations (1), (2), (5), and (6), the estimated coefficients for labor intensity are all positive, and significant in two cases, establishing a positive relationship between labor intensity and both nominal and effective *tariff* rates. However, when both tariffs and nontariff barriers are incorporated into a total measure of protection, all statistical significance drops away, so that equations (3) and (7) reveal no significant relationship, and equations (4) and (8) a negative but insignificant relationship between labor intensity and both nominal and effective total measures of restriction.

IV. Protection of Unskilled Labor

By considering that labor is not homogeneous, an interesting hypothesis to test is whether United States tariffs and nontariff barriers protect unskilled labor. Leontief (1956) has found that United States exports contain a significantly higher proportion of skilled workers than competitive import replacement, while Kravis (1956) confirmed that United States export industries generally pay higher wages than import-competing industries. These findings form the basis of an alternative resolution of the Leontief paradox. By allowing for human capital in factor intensities and heterogeneous labor, United States exports are skilled-labor intensive, and capital intensive when human capital is included, conforming to the Heckscher-Ohlin pattern of trade with unskilled labor being the "scarce" factor.

This led Ball (1967) to postulate the hypothesis that, based on the Stolper-Samuelson theorem, unskilled labor, the scarce factor of the United States, have an incentive to seek protection. His test of whether a positive relationship exists between protection and the proportion of unskilled labor in each industry is, however, unsatisfactory. From the *effective tariff rates* for the United States in 1962 provided by Balassa (1965), Ball chose 31 industries for which the average wage rate can be obtained. He found rank correlation coefficients of −.568 between the effective tariff rates and the average wage of all employees, and −.685 between the effective rates and the average wage of production workers. Next, he calculated the proportion of skilled workers (defined as professional, technical and kindred workers, and craftsmen and foremen) for the 20 two-digit SIC manufacturing industries for 1960 and compared these with the average wage for each industry in 1960. The rank correlation between average wage and the percentage of skilled workers is .94, while that between average wage of production workers and percentage of skilled workers is .932. On the basis of the high rank correlation between average wage and the proportion of skilled workers, Ball asserts that "average wage is a fair index of skill intensity," and since average wage is negatively related (in rank) to effective tariff rates, he concludes that "the 1965 United States tariff relatively heavily protected low wage United States industries using higher proportions of lower-skilled labor" (Ball 1967, p. 186).

There are three main objections to his procedure. In the first place, even though average wage is highly correlated with both effective rates and skill intensity, this does not necessarily imply that effective rates and skill intensity are highly correlated. Secondly, the correlation coefficients of .94 and .932 he found between average wage and skill intensity is for a higher level of industrial aggregation (20 two-digit industries) than that used in the correlation between average wage and effective tariff rates

TABLE 2

REGRESSION RESULTS: RATES OF PROTECTION ON PERCENTAGE OF UNSKILLED LABOR

DEPENDENT VARIABLE	INDEPENDENT VARIABLES Constant	 % Unskilled Labor	R^2
1 Nominal tariffs	5.80 (3.29) (1.76)	0.12* (0.07) (1.83)	0.06
2 Nominal tariffs and nontariff barriers	6.89 (3.78) (1.82)	0.12 (0.08) (1.55)	0.05
3 Effective tariffs	7.92 (8.91) (0.89)	0.25 (0.18) (1.38)	0.04
4 Effective tariffs and nontariff barriers	5.74 (9.91) (0.58)	0.31 (0.20) (1.52)	0.04

SOURCES.—Percentage of unskilled labor: unskilled workers defined to include operatives and kindred workers and laborers (constructed from U.S. Bureau of the Census 1963). Note that this is for 1960 (the closest year for which data are available) while the rates of protection are for 1964. The consequences of this lag can be expected to be minor: occupational composition remains relatively constant within a short span of time.

NOTE.—Estimated coefficients, followed by standard errors and *t*-statistics (estimated by ordinary least squares).

* Significant at 10% level.

(31 industries). The rank correlation coefficients may be sensitive to aggregation, and the results are therefore not strictly comparable. Finally, since the only protective rates included in his study are *effective tariff rates* with no consideration of nontariff barriers or nominal rates, his results are incomplete. If the intermediate products of a relatively low-skill industry are produced by an even higher proportion of unskilled labor, the effective rate of protection for the industry may be low even if its nominal rate is high.

To properly test the hypothesis that the United States protects more heavily industries that use a higher proportion of unskilled labor, we proceed to examine directly the relationship between the various measures of protection calculated from Baldwin and the percentage of workers classified as unskilled on the same level of industrial aggregation. The results of regressing rates of protection on the percentage of unskilled labor in each industry are presented in table 2. The estimated coefficients for the percentage of unskilled labor are all positive, but statistically significant at the 10 percent level only in equation (1) for *nominal tariff rates*. We can therefore conclude that while industries with a higher proportion of unskilled labor have higher nominal tariffs, they are not necessarily protected by effective rates or nontariff barriers.

V. Conclusions

Our empirical results confirm Vaccara's hypothesis that United State- nominal tariffs protect labor-intensive manufacturing industries. Hows ever, this finding on its own does not support Travis's argument that United States trade policy is responsible for the Leontief paradox. Any evidence which ignores nontariff barriers to trade and attempts to use only nominal and effective tariff rates in order to establish or refute Travis's hypothesis must be regarded as incomplete By including all restrictions, tariff and nontariff, in a total measure of protection and establishing that there is no statistically significant positive relationship between labor intensity and total measures of protection, our results therefore negate Travis's hypothesis

Finally, by taking into account variations in the skill mix of the labor force, we find that industries with higher proportions of unskilled labor are more heavily protected by nominal tariffs, but not by nontariff barriers or effective rates

References

Balassa, Bela "Tariff Protection in Industrial Countries An evaluation " *J P E* 73 (December 1965) 573–94

Baldwin, Robert E *Non-tariff Distortions of International Trade* Washington, Brookings Inst , 1970

Ball, David S "United States Effective Tariff and Labor's Share " *J.P E* 75 (April 1967) 183–87

Basevi, Giorgio "The United States Tariff Structure Estimation of Effective Rates of Protection of United States Industries and Industrial Labor " *Rev Econ and Statis* 48 (May 1966) 147–60

Cheh, John "United States Concessions in the Kennedy Round and Short-Run Labor Adjustment Costs " *J Internat Econ* 4 (1974) 323–40 (*a*)

——— 'United States Trade Policy and Short-Run Domestic Adjustments Costs " Ph D dissertation, Massachusetts Inst Tech , 1974 (*b*)

Kravis, Irving B "Wages and Foreign Trade " *Rev Econ and Statis* 38 (February 1956) 14–30

Leontief, Wassily "Factor Proportions and the Structure of American Trade Further Theoretical and Empirical Analysis " *Rev Econ and Statis* 38 (November 1956) 386–407

Ohlin, Goran "Trade in a Non-Laissez-Faire World " In *International Economic Relations* edited by Paul A Samuelson London Macmillan, 1969

Travis, William P *The Theory of Trade and Protection* Cambridge, Mass Harvard Univ Press, 1964

U S Bureau of the Census *U S Census of Population 1960 Occupation by Industry.* Washington Government Printing Office, 1963

U.S Bureau of the Census *Census of Manufactures 1963.* Washington Government Printing Office, 1966

Vaccara, Beatrice N *Employment and Output in Protected Manufacturing Industries* Washington Brookings Inst , 1960

Import Instability and Tariffs: Some Welfare Implications of Price Stabilization

Toshihisa Toyoda

Kobe University and North Carolina State University

I. Introduction

The welfare effects of price stability and/or instability have been discussed by many authors. Some 30 years ago, Waugh (1944) demonstrated that consumers could benefit from price instability. Later Oi (1961) showed this can also be the case for producers. Massel (1969) integrated these results and found that both consumers and producers could be better off by price stability if *actual compensation* was paid. Bieri and Schmitz (1974) considered the welfare effects of price instability when markets are noncompetitive, regulated by either a producer marketing board or a pure middleman.

Recently, Heuth and Schmitz (1972), and Bieri and Schmitz (B-S) (1973) extended the above analyses to international trading situations. Heuth and Schmitz, using a free trade model, showed that price instability caused outside a country's border can be preferable for both consumers and producers. B-S, using a model with monopoly power—either in the form of tariffs or marketing boards—rather than a free trade model, showed that international price stability is preferred for an importing country regardless of the source of instability when tariffs are in effect, but that this is not true when trade is restricted by export marketing boards rather than by tariffs.[1] This striking discrepancy between the two models of the welfare effects in an importing country has an important policy implication, in real trading situations, it is a very common practice to impose tariffs, and therefore stabilizing prices by buffer stocks must be preferable for an importing country so long as the B-S conclusion is correct. This is particularly so for agricultural products, with which most of the authors in the literature of welfare effects of price stabilization are concerned, and to which this type of analysis can safely be applied.

My special thanks to an unknown referee, whose penetrating comments have made my analysis clear and compact. I also acknowledge the helpful comments of Paul R. Johnson, Martin Bronfenbrenner, Tom J. Grennes, and Marie Thursby.

[1] B-S (1974) also showed that an exporting country can prefer instability regardless of the type of monopoly power in trade if the source of instability is external.

[*Journal of Political Economy*, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved

In this note I show that if the importing country does not have complete monopsony power, in contrast to the B-S assumption, it does not necessarily prefer price stabilization even if tariffs are imposed I consider changes in the supply conditions in the exporting country as the only sources of instability, it is reasonable to disregard changing demand situations in the importing country not only because it simplifies my analysis but also because the demand for agricultural products is fairly stable compared with the frequent shifts in supply I also introduce a new notion of "optimal" tariffs related to price stability.

II. Analysis of Welfare Effects

Throughout the analysis I assume perfect competition in production in the importing country as well as in the supply from the exporting country

We can assume various alternative behaviors of the government of the importing country B-S (1973) implicitly assume that the government has complete monopsony power and that it maximizes producers' returns, that is, both profits and producers' surplus Even if the government has complete monopsony power, it can also be assumed to maximize profits only, that is, it sets the marginal cost equal to the marginal revenue expressed in the domestic price However, I feel the complete monopsony assumption is very awkward and unrealistic and reject its use for the following reasons First, in real agricultural trades tariffs are usually set only for the purpose of protecting domestic suppliers and may be below the "optimal" levels in the usual partial equilibrium sense Second, it is contradictory to assume complete monopsony power since the importing country's government on the one hand is assumed to *know* exactly the changing marginal costs and therefore the changing foreign supply conditions, but on the other hand is assumed to assign equal probabilities for possible foreign supply conditions because it *does not know* them In the following discussion, instead of assuming that the government has complete monopsony power and sets different optimal tariff rates for any shifts in foreign excess supply (as in the case of B-S), I assume that in this partial equilibrium framework the tariff rate is given exogenously and is constant for any exogenous shifts in foreign excess supply

Let me set up the model for the case in which a tariff is imposed but no buffer stocks are created by the government Let us consider two situations corresponding to two fluctuating excess supplies and their prices Without loss of compatibility with the geometric analysis by B-S (1973), I consider the following linear model [2]

---

[2] If a quota is imposed instead of a tariff and it is allocated so as to ensure perfect competition among quota holders, the following analysis can equally be applied (see, e.g , Bhagwati [1965]).

Situation 1		Situation 2	
$ED_{M1} = \alpha_0 - \alpha_1 P_{M1}$	(1)	$ED_{M2} = \alpha_0 - \alpha_1 P_{M2}$	(1′)
$ES_{x1} = -\beta_0 + \beta_1 P_{x1}$	(2)	$ES_{x2} = -\beta_0' + \beta_1 P_{x2}$	(2′)
$P_{x1}(1 + t) = P_{M1}$	(3)	$P_{x2}(1 + t) = P_{M2}$	(3′)
$ES_{x1} = D_{M1}$	(4)	$ES_{x2} = D_{M2}$	(4′)

$$\beta_0 \leq \beta_0' \quad (5)$$

where α_i, $\beta_i (i = 0, 1)$, and β_0' are all nonnegative constants Equation (1) states that the domestic excess demand is a function of domestic price, equation (2) that the foreign excess supply is a function of foreign price, equation (3) that the domestic price exceeds the foreign price by the amount of the tariff, and equation (4) that the foreign excess supply equals domestic excess demand Equations (1′)–(4′) correspond to definitions (1)–(4) above except in the excess supply function from the exporting country (2′), the intercept can shift subject to (5) [3]

For each situation we have four equations and five unknowns ED_M, ES_x, P_M, P_x, and t. If t is given, the remaining unknowns are determined Consequently, corresponding to every t, there will be some levels of quantities and prices

Now, assume the government of the importing country assigns equal probabilities to P_{M1} and P_{M2} and stabilizes them at $P_u = (P_{M1} + P_{M2})/2$ through a costless storage activity [4] As has been shown by Heuth and Schmitz (1972), the net social loss (S L) for consumers and producers due to price stabilization is depicted in figure 1 by $maP_{M1}P_u - dmP_uP_{M2}$ It can be expressed algebraically as

$$\begin{aligned} \text{S L} &= \alpha_1 (P_u - P_{M1})^2 \\ &= \alpha_1 \{(\beta_0' - \beta_0)(1 + t)/2[\alpha_1(1 + t) + \beta_1]\}^2 \geq 0 \end{aligned} \tag{6}$$

The government's tariff revenue can generally be expressed as $[t/(1 + t)]P_M ES_x$ Therefore, the difference between the gains and losses in tariff revenues due to price stabilization (R G), that is, $jklP_u$ +

---

[3] We have also analyzed the situation in which β_0 and/or β_1 may shift and obtained essentially the same results In this case, however, the market is not necessarily cleared by the government's buffer stock policy and, as the referee pointed out, the government's expected gain from stabilization should be taken into account for a further investigation

[4] Introducing a proportional storage cost function needs only minor modifications of algebra and does not change the main conclusions below

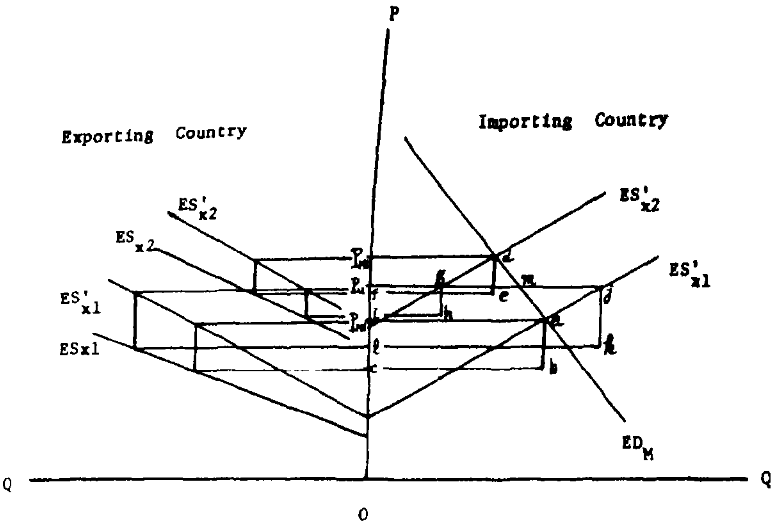


FIG 1.—Welfare effects with tariffs

$ghiP_u - abcP_{M1} - defP_{M2}$ in figure 1, can be expressed as

$$\begin{aligned} \mathrm{RG} &= \left(\frac{t}{1+t}\right)(P_u - P_{M1})\left[-\beta_0 + \frac{\beta_1}{1+t}P_{M1} - \left(\beta_0' + \frac{\beta_1}{1+t}P_{M2}\right)\right] \\ &= \left(\frac{t}{1+t}\right)(P_u - P_{M1})(ES_{x1}' - ES_{x2}') \qquad (7) \\ &= \alpha_1(\beta_0' - \beta_0)^2 t(1+t)/\{2[\alpha_1(1+t)+\beta_1]^2\} \geq 0, \end{aligned}$$

where ES_{x1}' and ES_{x2}' are foreign excess supplies, with tariffs expressed in terms of domestic prices, for situations 1 and 2, respectively It is clear from (7) that the tariff revenue under price stability is greater than the tariff revenue generated from instability if a tariff is imposed and a shift in the excess supply function occurs, this is the B-S result

Next, let us compare the social losses expressed by (6) with the revenue gains expressed by (7) Subtracting (6) from (7), we get the net gain,

$$\mathrm{NG} = -[\alpha_1(\beta_0' - \beta_0)^2(1 - t^2)]/\{4[\alpha_1(1+t)+\beta_1]^2\}, \qquad (8)$$

which can be negative in contradiction to the B-S conclusion cited in the previous section The sign of (8) critically depends on the value of t Assuming $\beta_0' > \beta_0$, it is seen that the net gain after an actual compensation payment from the government to consumers and producers becomes positive, zero, and negative according to:

$$t \gtreqless 1 \qquad (9)$$

Therefore, $t = 1$ is the critical value of the tariff rate for discriminating positive and negative net gains in the importing country. It should be noted that this critical point does not depend on any underlying parameters, although it does depend on the assumption of linearity.

What is the "optimal" tariff policy in this situation? Here, the optimal tariff is defined as the one which maximizes (8), the net gain in the importing country attained after the government has stabilized prices and made an actual compensation to consumers and producers. This is quite a different notion of the usual optimal tariff which does not take price stabilization into consideration. Differentiating (8), we find

$$d(\mathrm{N\,G})/dt > 0, \text{ for } t \geq 0 \qquad (10)$$

That is, the net gain is a monotone increasing function of $t(\geq 0)$ and therefore a unique maximum cannot be attained, the larger the value of t, the greater the net gain Therefore, no optimal tariff policy exists in this case, provided that there is no tariff redundancy

The above analysis shows that the importing country can benefit net positive welfare only by imposing more than 100 percent tariff rates Therefore, under usual tariff rates of less than 100 percent, the importing country prefers price instability to stability Note that I have assumed no prior knowledge of the foreign excess supply conditions, in contrast to the B-S case If prior knowledge of them is available, we can of course attain the usual unique optimal tariffs corresponding to given foreign supply situations Also note that our analysis above assumes a competitive foreign supply and therefore that there are no governmental interventions or counterpolicies on the export side [5]

III. Conclusion

With particular reference to trade in agricultural products, I have analyzed some welfare effects of price stabilization by the government of an importing country In contrast to the B-S analysis, I have assumed that the importing country has no complete monopsony power tariffs are imposed only for the purpose of protecting a small domestic supply and the tariff rate is given exogenously

The net gain of the importing country, after the governmental compensation payment, is positive if the tariff rate is greater than 100 percent However, if there is no tariff redundancy the optimal tariff rate does not exist, since the net welfare gain is a monotone increasing function

---

[5] Assuming that the government of the exporting country has complete monopoly power and sets the marginal revenue equal to the marginal cost, we can also show algebraically that the importing country prefers price instability to stability

of tariff rate Under usual circumstances, with less than 100 percent tariff rates, the importing country loses in net welfare by stabilization and therefore prefers price instability to stability

References

Bhagwati, J "On the Equivalence of Tariffs and Quotas." In *Trade, Growth, and the Balance of Payments*, edited by H G Johnson and P B. Kenen. Chicago: Rand-McNally, 1965

Bieri, J, and Schmitz, A "Export Instability, Monopoly Power and Welfare," *J Internat Econ* 3 (November 1973) 389–96

——— "Market Intermediaries and Price Instability Some Welfare Implications " *American J Agricultural Econ* 56 (May 1974) 280–85

Hueth, D, and Schmitz, A "International Trade in Intermediate and Final Goods Some Welfare Implications of Destabilized Prices " *Q.J.E* 86 (August 1972) 351–65

Massel, B F "Price Stabilization and Welfare " *Q J E* 83 (May 1969) 285–97.

Oi, W Y "The Desirability of Price Instability under Perfect Competition " *Econometrica* 29 (January 1961) 58–64

Waugh, F V "Does the Consumer Benefit from Price Instability?" *Q.J E* 58 (August 1944) 602–14

The Leff Conjecture: Some Contrary Evidence

John R. Hanson II

Texas A & M University

In a recent issue of this *Journal*, Leff (1973) attributes the failure of export-led development to occur in nineteenth-century Brazil to the slow growth of exports, especially in per capita terms, and to the small share of exports in Brazilian national product Leff speculates that during the same period other less developed countries (LDCs) failed to garner dynamic gains from trade for the same reasons This conjecture contrasts with most of the existing literature, which treats imperfections in the linkages between internal development and the external trade stimulus as a main reason for persistent poverty in the Third World

In Section I, I present estimates of per capita exports in a large sample of LDCs for the years 1860, 1880, 1900, and 1913 (table 1) and of growth rates of per capita exports in these countries during two intervals of time, 1860–1900 and 1900–1913 (table 2) In Section II, I use the results to test Leff's conjecture The periodization used here does not correspond exactly to the 1850–96 and 1897–1913 periods on which Leff focuses, in part because a sufficient quantity of data going back to 1850 is not available Also, Leff's periodization of the nineteenth century is based on residuals from trend equations estimated for a particular country, Brazil, and may not be applicable to other LDCs It is sufficient that the estimates span a long period of time, be widely spaced chronologically, and permit subdivision of the whole period into reasonably homogeneous segments For purposes of brevity and in keeping with the spirit of Leff's article, the analytical discussion focuses on the period 1860-1900

I. Data

Most of the export data are taken from a remarkable but apparently little-known source, the *Statistical Abstract of Foreign Countries*, published in 1909 by the United States Department of Commerce and Labor The *Abstract* contains time series of export values for 60 countries dating back

[*Journal of Political Economy*, 1976, vol 84, no 2]
© 1976 by The University of Chicago All rights reserved

in some cases to the first half of the nineteenth century. These time series were carefully compiled from the official trade reports of the countries surveyed, and communication with foreign governments was undertaken to clear up occasional problems, such as the proper conversion ratios of foreign currencies into United States dollars Since the *Abstract* is not comprehensive, Hanson (1972), a statistical survey of nineteenth-century LDC trade, has been used to supplement it For the year 1913, the export data are taken from Yates (1959) All data are f o b

Most of the population data are taken from the list given in Zimmerman (1962) These data, which pertain only to Latin America and Asia, were compiled by Zimmerman and his associates from various sources and stand up well to consistency checks For example, Zimmerman places Colombia's population in 1860 at 2 3 million people, which is close to the 2 million figure for 1850 given by McGreevey (1971) Furthermore, totals of Zimmerman's disaggregated estimates for Asia and Latin America are within the margins of error of the aggregate estimates for these regions presented in Durand (1966) However, for China's population I used Durand's estimates and for British Guiana, Egypt, and Jamaica, which are not listed by Zimmerman, other secondary sources listed below [1]

Because of the lack of reliable population data for most of Africa, only one country from this region, Egypt, is included in table 1 However, Durand gives aggregate population estimates for sub-Saharan Africa in the nineteenth century, and, in order to gain at least a rough impression of the degree of involvement of Tropical Africa in trade during this period, I have used his "medium" estimates in conjunction with the comprehensive export data in Hanson (1972) to calculate per capita exports for that region in 1860 and 1900 The results of these calculations are presented in table 1 but are not included in subsequent statistical comparisons, even though, taken at face value, they would strengthen the argument

Whenever possible, 3-year averages of the export data centered on the benchmark years were used in computing the per capita figures When a single-year export figure was the only one available, the estimate utilizing that figure is denoted in table 1 by parentheses I have not followed Leff in deflating the export data to estimate long-term trends in the income terms of trade Price indexes of imports are available for too few tropical countries to make this procedure feasible, and it would be pointless to adjust most or all of the countries surveyed by the Imlah index of British export prices, as Leff does for Brazil, since international comparisons would not be affected Also, presenting the data in unadjusted form gives the estimates greater flexibility for other scholarly purposes

[1] See Mandle (1973), Issawi (1961), and Eisner (1961) Also, the per capita export figures for Ceylon and Colombia are taken directly from Lim (1968) and McGreevey (1971)

TABLE 1

PER CAPITA EXPORTS, SELECTED COUNTRIES, 1860–1913
(3-YEAR AVERAGES, IN DOLLARS)

	1860	1880	1900	1913
Central America				
Costa Rica	(15.00)	n.a.	25.00	(33.33)
Cuba	(19.07)	n.a.	33.13	(65.55)
Guatemala	(1.90)	n.a.	(7.50)	(7.78)
Honduras	(1.50)	n.a.	(11.80)	(8.34)
Jamaica	12.22	10.69	13.54	(12.63)
Mexico	n.a.	2.59	2.54	(5.34)
Nicaragua	(0.75)	n.a.	5.60	(13.34)
Puerto Rico	(9.16)	n.a.	(14.11)	n.a.
South America				
Brazil	5.90	7.17	8.94	(12.52)
British Guiana	46.00	51.25	27.03	n.a.
Chile	13.56	22.17	21.03	(41.43)
Colombia	3.60	8.40	4.10	(5.80)
Ecuador	(3.00)	(4.63)	8.56	(16.00)
Peru	(9.96)	(7.14)	5.66	(9.77)
Uruguay	17.80	27.61	35.89	(65.45)
Venezuela	3.38	7.00	8.74	(12.18)
Asia				
British India	0.63	1.06	1.26	(2.73)
Ceylon	7.78	5.83	9.72	(18.95)
China	0.18	0.24	0.29	n.a.
Indonesia	1.77	2.47	2.53	(5.61)
Philippines	(2.00)	3.62	2.67	(5.74)
Thailand	1.00	1.32	2.24	(5.18)
Africa				
Egypt	(3.49)	10.00	8.12	(13.08)
Tropical Africa	(0.33)	n.a.	(1.80)	n.a.
Mean (excluding Tropical Africa)	8.17	10.19	11.30	(18.04)
Median (excluding Tropical Africa)	3.60	7.00	8.74	(12.52)

SOURCES.—See text.

NOTE.—Figure for Cuba in col. 3 is 3-year average centered on 1901; for Colombia, annual averages, 1855–59, 1875–79, 1895–99; for Uruguay, col. 1 figure is 3-year average centered on 1864; for Venezuela, 3-year averages centered on 1854, 1876, and 1894; for Ceylon, annual averages for decades 1860–69, 1880–89, 1900–1909, 1910–19; for China, first figure is 3-year average centered on 1865; for Indonesia, 1860 export data come from Furnivall (1939, p. 170); and for Thailand first figure is 3-year average centered on 1869 while 1913 export data come from Maddison (1971, p. 59). Parentheses mean estimate is for year at column head only.

II. Brazil in International Perspective

On one reading, the evidence in tables 1 and 2 is inconsistent with Leff's conjecture. Taking the level of per capita exports as a proxy for the share of exports in national product, Brazil was merely an "average" LDC according to table 1. Per capita exports in Brazil were at or above the median of the sample in 1860, 1880, and 1900, but below the mean at each of those dates. According to table 2, the Brazilian rate of per capita export growth was below the LDC average no matter whether the mean or median is taken as the standard of comparison.

However, if table 1 is read in relation to table 2, an interpretation more favorable to Leff is possible. Only three countries with high per capita

TABLE 2

ANNUAL AVERAGE GROWTH RATE OF PER CAPITA EXPORTS, SELECTED COUNTRIES, 1860–1900 AND 1900–1913

	1860–1900	1900–1913
Central America		
Costa Rica	1.3	2.2
Cuba	1.4	5.9
Guatemala	3.5	0.3
Honduras	5.3	−2.5
Jamaica	0.3	−0.5
Mexico	n.a.	5.9
Nicaragua	5.1	6.9
Puerto Rico	1.1	n.a.
South America		
Brazil	1.1	2.4
British Guiana	−1.3	n.a.
Chile	1.1	5.4
Colombia	0.3	2.5
Ecuador	2.6	4.9
Peru	−1.4	4.3
Uruguay	2.0	4.8
Venezuela	2.4	1.8
Asia		
British India	1.8	6.2
Ceylon	0.6	6.9
China	1.2	n.a.
Indonesia	0.9	6.3
Philippines	0.7	6.1
Thailand	2.0	6.2
Africa		
Egypt	2.1	3.7
Mean	1.6	4.0
Median	1.3	4.9

SOURCE.—Table 1.

exports in 1860 subsequently achieved a rate of per capita export growth higher than Brazil's. These are Costa Rica, Cuba, and Uruguay. British Guiana and Peru experienced a decline in per capita exports, and in Jamaica per capita exports were virtually stagnant. On the other hand, several countries, such as Honduras and Nicaragua, experienced rapid export growth but began the period with low exports per capita. The same is probably true of many countries in Tropical Africa.[2] For the complete sample of LDCs (excluding Tropical Africa) the Spearman coefficient of rank correlation between the rate of per capita export growth between 1860 and 1900 and the level of per capita exports in 1860 is −0.48, which is significant above the 0.05 level.

The point may be stated summarily. An analysis of growth failure that is based on a low rate of per capita export growth *and* a small export sector is probably not applicable to most tropical countries before 1900 It *is* likely, however, that most LDCs had at least one of these characteristics Even this weaker statement is sufficient to establish the presumption that, regardless of internal conditions, significant economic development would have been difficult if not impossible for the less developed countries to achieve through expanding exports in the second half of the nineteenth century.

---

[2] The sugar islands of Reunion and Mauritius probably are exceptions to this statement. These islands were similar to parts of the Caribbean in many ways and probably had high ratios of exports to national product and low rates of per capita export growth.

References

Durand, John "The Modern Expansion of World Population." *Proc. American Philosophical Soc.* 110 (November 1966) 136–59

Eisner, Gisela *Jamaica, 1830–1930, A Study in Economic Growth* Manchester. Manchester Univ Press, 1961.

Furnivall, John S *Netherlands India* London Cambridge Univ Press, 1939

Hanson, John R., II *The 19th Century Exports of the Less Developed Countries.* Ann Arbor, Mich Univ Microfilms, 1972

Issawi, Charles "Egypt Since 1800 A Study in Lop-sided Development" *J Econ History* 21 (March 1961) 1–26

Leff, Nathaniel "Tropical Trade and Development in the Nineteenth Century The Brazilian Experience" *J P E* 81, no 3 (May/June 1973) 678–96

Lim, Youngil "Trade and Development The Case of Ceylon" *Econ Development and Cultural Change* 16, no 2 (January 1968) 245–60

McGreevey, William P *An Economic History of Columbia, 1845–1930* London Cambridge Univ. Press, 1971

Maddison, Angus *Class Structure and Economic Growth* New York Norton, 1971

Mandle, Jay R *The Plantation Economy* Philadelphia Temple Univ Press, 1973

U S Department of Commerce and Labor *Statistical Abstract of Foreign Countries* Washington Government Printing Office, 1909

Yates, P Lamartine *Forty Years of Foreign Trade* New York Macmillan, 1959

Zimmerman, L J "The Distribution of World Income, 1860–1960" In *Essays on Unbalanced Growth*, edited by E De Vries The Hague Mouton, 1962

A Note on Risk, Return, and Equilibrium: Empirical Tests

Menachem Brenner

The Hebrew University, Jerusalem

In their rigorous empirical testing of the two parameter model, Fama and MacBeth (FM) (1973) use the cross-sectional regression

$$R_{pt} = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}\hat{\beta}_{p,t-1} + \hat{\eta}_{pt} \tag{1}$$

for each month t of their "testing period," where R_{pt} and $\hat{\beta}_{p,t-1}$ are, respectively, the average rate of return and systematic risk across securities in portfolios The time series coefficients of (1) are used for testing the two parameter model, where "the least-squares estimate $\hat{\gamma}_{0t}$ can always be interpreted as the return for month t on a zero-$\hat{\beta}$ portfolio" (FM 1973, p 626) [1]

In this note we test the empirical validity of this claim The results of this test may have implications for some tests conducted by FM and for tests of capital market efficiency that use $\hat{\gamma}_{0t}$ as a second market factor In the context of "market models," the coefficient $\hat{\gamma}_{0t}$ is interpreted as a market factor (in addition to R_{mt}) that makes up the "two factor market model" [2]

$$R_{pt} = \hat{\gamma}_{0t}(1 - \hat{\beta}_p) + \hat{\beta}_p R_{mt} + \hat{\eta}_{pt} \tag{2}$$

This model became increasingly popular among researchers that test market efficiency (e g , Ball 1972, Jaffe 1974, Mandelker 1974)

To test $\hat{\beta}_0 = 0$ we use the estimate

$$\hat{\beta}_0 = \frac{\sum_t (\hat{\gamma}_{0t} - \hat{\gamma}_0)(R_{mt} - \bar{R}_m)}{\sum_t (R_{mt} - \bar{R}_m)^2}, \tag{3}$$

where the $\hat{\gamma}_{0t}$'s are obtained from (1) [3] and the R_{mt}'s are the Fisher Arithmetic Index returns This estimate was computed for the entire

I am indebted to Amir Barnea, Dan Galai, Seymour Smidt, and an anonymous referee for their helpful comments The comments of Dan Galai and the anonymous referee have substantially improved the relevance of this note

[1] In n 10 FM explain that the $\hat{\beta}_0$ property of $\hat{\gamma}_{0t}$ follows from the Gauss-Markov Theorem

[2] FM arrive at (2) using the regression property $\hat{\gamma}_{0t} = \bar{R}_{pt} - \hat{\beta}_p\hat{\gamma}_{1t}$, but, $\bar{R}_{pt} = R_{mt}$ and $\hat{\beta}_p = 1$ Thus, $\hat{\gamma}_{1t} = R_{mt} - \hat{\gamma}_{0t}$, which they substitute in (1) to obtain (2)

[3] The $\hat{\gamma}_{0t}$'s were kindly provided by Fama and MacBeth

[*Journal of Political Economy*, 1976, vol 84, no 2]
© 1976 by The University of Chicago All rights reserved

TABLE 1

$\hat{\beta}_0$ FOR THE FM PERIODS

	1935–6/68	1935–45	1946–55	1956–6/68	1935–40	1941–45	1946–50	1951–55	1956–60	1961–6/68
$\hat{\beta}_0$	139	113	238	168	134	026	276	113	075	177
$t(\hat{\beta}_0)$	4 59	2 23	4 76	2 77	1 94	0 33	4 03	1 61	2 53	2 16

period January 1935–June 1968 and for all subperiods tested by FM Table 1 provides the values of $\hat{\beta}_0$ and the corresponding t statistics[4] The results in table 1 suggest that we reject the hypothesis $\hat{\beta}_0 = 0$ $\hat{\beta}_0$ is positive in all periods, and eight out of 10 estimates are significant The t values are large for the overall period (4 59) and for all three subperiods in the long subperiod breakdown (2 23, 4 76, 2 77) Though significantly different from zero, the size of $\hat{\beta}_0$ is rather small (in the overall period $\hat{\beta}_0 = 139$ compared with $\beta = 1$ for the average security)

These results have some implications for the FM tests The tests on $\hat{\gamma}_1 = R_m - \hat{\gamma}_0 > 0$ are strengthened A $\hat{\beta}_0 >$ makes $R_m - \hat{\gamma}_0$ biased downward Therefore, the hypothesis of positive expected return-risk tradeoff is supported more strongly than if $\hat{\gamma}_{0t}$ were truly the return on a zero-β portfolio The results of the tests of the Sharpe-Lintner hypothesis using $\bar{\gamma}_0 = R_f$ are somewhat biased upward A more appropriate test would be $\hat{\gamma}_0 - \hat{\beta}_0 R_m = R_f$, where the bias due to $\hat{\beta}_0 > 0$ is corrected This correction would reduce the deviation of $\hat{\gamma}_0$ from R_f, *but* this reduction may be small due to the small value of $\hat{\beta}_0$ Thus the main conclusions about the Sharpe-Lintner hypothesis may not be seriously affected

A more important consequence of these results is the bias introduced in tests of market efficiency when the "two-factor market model" (2) is used For $\hat{\beta}_p < 1$ the residual ($\hat{\eta}_{pt} = R_{pt} - \hat{R}_{pt}$) is biased downward, while for $\hat{\beta}_p > 1$ the residual is biased upward The size of this bias is a function of $|1 - \hat{\beta}_0|$ and it may seriously affect these tests

Finally, the results here are interesting for what they imply about the way $\hat{\beta}_p$ change through time In the FM tests $\hat{\gamma}_{0t}$ is the return on a portfolio constructed from the 20 portfolios used by FM The $\hat{\beta}_p$ of the 20 portfolios are weighted in such a way that $\hat{\gamma}_{0t}$ has $\hat{\beta}_0 = 0$ (by the Gauss Markov Theorem) But, the $\hat{\beta}_p$ are based on data for period $t - 1$ and not period t Thus $\hat{\gamma}_{0t}$ is the return on a $\hat{\beta}_0 = 0$ portfolio only if $\hat{\beta}_p$ remain unchanged from period to period The results here indicate that $\hat{\beta}_p$ change in such a way that when the $\hat{\beta}_0$ of $\hat{\gamma}_{0t}$ is computed it turns out to be slightly but consistently positive. The implied change in $\hat{\beta}_p$ is of the tendency toward one nature. For $\beta_p > 1$, $\hat{\beta}_{pt} < \hat{\beta}_{pt-1}$, and for $\beta_p < 1$, $\hat{\beta}_{pt} > \hat{\beta}_{pt-1}$ [5]

---

[4] An assumption of stationarity is implicit in the estimation and interpretation of these statistics

The implications of the finding that $\hat{\beta}_0$ of $\hat{\gamma}_{0t}$ is not zero can be summarized as follows: (1) Has a limited effect on the FM tests; (2) may be of importance for testing market efficiency, (3) $\hat{\beta}_p$ tend, over time, toward the average value of 1.0

References

Ball, R. "Changes in Accounting Techniques and Stock Prices." Ph.D dissertation, Univ Chicago, 1972.

Blume, M E. "On the Assessment of Risk " *J Finance* 26 (March 1971): 1–10.

Fama, E F , and Macbeth, J D "Risk, Return and Equilibrium· Empirical Tests " *J P E.* 81 (May-June 1973) 607–36

Jaffe, J F "Special Information and Insider Trading " *J. Bus.* 47 (July 1974)· 410–28

Mandelker, G "Risk and Return The Case of Merging Firm " *J Financial Econ* 1 (December 1974) 303–35

---

[5] This corroborates similar findings of this nature in the past (Blume 1971)

Miscellany

Toward a Deeper Economics of Sleeping

Exhaustive study of Professor M. A. El Hodiri's tedious, yet curiously superficial analysis of the economics of sleeping has led to a powerful and remarkably boring reformulation of Hodiri theory.[1]

As Hodiri astutely observes, it is difficult to deny that a sensible man would seek to maximize $U(x, y) = x^2y$ where x is daily consumption and y is the fraction of the day spent in bed. Hodiri supposes that daily consumption equals daily income, which in turn equals the daily wage rate, w, times the fraction of the day one spends out of bed.[2] Hence $x = w(1 - y)$. But then $U = w^2(1 - y)^2y$. By substituting all real numbers between zero and one into this expression, the reader will quickly verify that U is maximized when $y = \frac{1}{3}$. Thus Hodiri asserts that, regardless of the wage rate he receives, a sensible man will sleep $\frac{1}{3} \times 24 = 8$ hours per day.

Had Hodiri been more deeply embedded in the profundities of economic literature, he could hardly have closed his eyes to the abundance of platitudinous evidence that some reap who do not sow. This fertile observation makes two things obvious. One is that the Hodiri equation, $x = w(1 - y)$, must be replaced by $x = w[(1 - y) + (R/w)]$ where R is daily earnings from nonlabor sources. The proper substitution for $U = x^2y$ is then $U = w^2[1 - y + (R/w)]^2y$. Consequently, U is maximized when $y = \frac{1}{3}[1 + (R/w)]$. This inescapable truth would appear to rend Hodiri theory all to pieces. How can it be that sensible men have

---

[1] It would be difficult to underestimate the importance of El Hodiri's contribution in enabling the author to achieve these heights of pedantry. Historians of thought could no doubt find many wearisome parallels in the advance of science. Though I do not presume to have mounted the shoulders of a giant, I hope that I have trod gently on the toes of El Hodiri.

[2] Professor J. T. Little of Washington University, St. Louis, in an unpublished note, maintains that not all time out of bed is spent in productive labor. However, there is reason to suppose that Little's observations have been confined to a pathological sample of the labor force.

[Journal of Political Economy, 1976, vol. 84, no. 2]
© 1976 by The University of Chicago. All rights reserved.

always slept 8 hours unless R/w has always been both constant and zero?

Resolution of this fearsome paradox will display once again to the jaded reader the beauty and the mystery of Economic Fairyland. Observe that the source of nonwage income is called "capital." Where r is the daily interest rate, K is the stock of capital, and N is the size of the labor force, the average daily nonwage income is $R = rK/N$. Then $R/w = (1 - y) rK/(1 - y)wN$. Since total daily labor income is $(1 - y)wN$, $rK/(1 - y)wN$ is the ratio of capital's share to labor's share of national income, which is well known to have been remarkably constant over time and is in fact equal to $\frac{1}{4}$. Hence in all historical epochs, the average value of R/w has been $(1 - y)\frac{1}{4}$. Recalling that $y = \frac{1}{3}[1 + (R/w)]$, making the appropriate substitution, and solving for y, we have $y = \frac{5}{13}$. Now $\frac{5}{13} \times 24 = 9.231$ hours. Thus have we elegantly vindicated Hodiri's tiresome "first law of soporifics," namely, "In all epochs the average man has spent the same amount of time in bed."

There remains a yawning chasm between Hodiri's banal "second law of soporifics" which is "most people sleep 8 hours a day" and our unassailable demonstration that the average man has always spent 9.231 hours per day in bed. As in the science of astronomy, the existence of the planet Pluto was first conjectured as a consequence of a remarkable deviation of planetary orbits from those predicted by pure theory, so must we conjecture that the average man spends 1.231 hours per day in bed doing something other than sleeping. I eagerly await the announcement that some penetrating theorist or tireless data-massager has actually discovered Activity X.

T. C. BERGSTROM

Washington University

Reference

El Hodiri, M. A. "The Economics of Sleeping." Unpublished manuscript, University of Kansas, 1973.

Book Reviews

Equilibrium and Disequilibrium. A Quantitative Analysis of Economic Interaction. By
MICHAEL ALLINGHAM
Cambridge, Mass. Ballinger Publishing Co, 1973. Pp xi + 161 $15.00

This book is not quite what it seems Its first part consists of a theoretical general equilibrium model, which is claimed to embody "a fundamental reformulation of the concepts of station and movement—or equilibrium and disequilibrium" (p 3). This reformulation is then applied to a model of an actual economy, and finally the empirical results are surveyed and appraised However, it soon becomes apparent, and is confirmed in the preface, that the econometric model came first and that the theory which it illustrates was developed later Of course this is not necessarily an invalid way of going about economic research, but in this case neither are the theoretical innovations of the first part very startling nor is there much relation between them and the model of the second part In fact any econometrician, on examining that model, would be surprised indeed to learn that it represented a fundamental reformulation of anything, since it is a very respectable piece of conventional empirical work

Consider first the basis for Allingham's dissatisfaction with standard general equilibrium theory. There are two main charges The first is that equilibrium is defined independently of any notion of stability He claims (p 11) that in fact "one cannot validly even define equilibrium in an economy without at the same time defining stability " This is obviously wrong, especially as "defining stability" means specifying the actions of agents in all circumstances The great attraction and comparative simplicity of equilibrium analysis is that it is logically self-consistent without having to ask what happens out of equilibrium The real target should be the hopeless inadequacy of Walrasian stability analysis Until it is possible to do without *tâtonnement*, or recontracting, there is no chance of ever relating theoretical models to the interesting and crucially important disequilibrium problems of the real world Sadly, *tâtonnement* is one aspect of conventional theory that Allingham cannot do without

The second objection to standard theory is that it assumes perfect competition, although the complaint is never stated in that precise form It appears when it is assumed that the variables which describe an agent's environment need not be exclusively price signals and that his choices do not consist entirely of quantities supplied and demanded It appears too when he condemns the Walrasian *tâtonnement* process for the artificial device of an auctioneer to change prices, although the force of this criticism is lessened by the inadequacy of the *tâtonnement* process itself There are, of course, well-known problems in introducing noncompetitive behavior into equilibrium analysis, some of which are not considered For example it appears (p 12) that while an agent can observe, say, the actual quantities demanded by other agents he is allowed no idea of their demand functions—nor is there any scope for learning It is apparent that Allingham throughout has in mind a world where neither information nor computational ability is complete; in that case he needs to be rather more careful about his assertion that in his model "the formation of equilibrium is completely explained by the independent optimising behavior of individual agents" (p 7) if he wants to say more than that it must be optimizing because it is what they do. So although the approach is more general than Walrasian *tâtonnement* it does not seem to contribute very much to the two central problems of disequilibrium trading and noncompetitive behavior

However the core of the book is the econometric model This is a quarterly model of the United Kingdom from 1956 to 1966 Production is divided into ten "industries," consumption into four categories, and traded goods into two, so that it is reasonably disaggregated The postulated functional forms are all quite simple, as all nonlinearities arise as ratios The various equations for consumption, investment, inventories, and so on take little account of the sophisticated econometric work that has been applied to them However, "its main purpose is to present a tested methodology for this type of research, rather than to answer more specific questions" (p 131), so that one should identify and evaluate its departures from conventional theory and empirical models On the theoretical side, the model has nothing to say about movement towards an equilibrium, it assumes (short-period) equilibrium to hold in every period, as indeed it must since by hypothesis any deviations are notional and subject to recontract, and so would not be observed As for relaxation of perfect competition, it is certainly true that variables other than prices appear on the right-hand side of the equations, and to the extent that one can regard the equations as legitimate aggregation of individuals' decision functions rather than equilibrium conditions, then that claim is indeed met in the model On the other hand it is a little ambitious to expect that "the" price of chemicals is set by the chemical industry—and that is one of the more homogeneous sections It is perfectly reasonable to assume that an aggregate mark-up relation for price is a good predictor, but not that it is a behavioral relation Finally it is meaningless to assert that the equations represent the "independent optimising behavior of individuals" since that is neither tested nor testable

Thus there is little connection between the theory of the first part of the book and the model of the second, but the model does have some interesting features of its own First it is quite large There are 79 stochastic equations, which are estimated by two-stage least squares Second, it pays commendable attention to supply considerations, both of capital and labor, in each industry However this leads to some disturbing features For one thing it is assumed that labor supply is entirely endogenous, and further that current labor produces current output and current labor, so that in principle there is no limit to employment and hence output, especially as there is no balance of payments constraint either This suggests that the model would be of limited use in policy making Furthermore, there is some confusion about the precise form of the production function On page 63 it is said to be linear, while on page 64 it is assumed that "proportional utilisation of labour force and capital stock are equal," which is a little unlikely if the elasticity of substitution is infinite There is also a difficulty about the capital stock term, for he uses a series which makes one assumption about depreciation, namely a fixed life for each type of asset, and then in the investment equations makes another, namely proportional depreciation But given that the series all look very like time trends this is unlikely to matter much

Overall, though, most of the estimated equations turn out with reasonable coefficients and seem to fit pretty well, and there is no doubt that the model is rather better than some to have appeared in the last few years. But it does not represent a fundamental reformulation, nor does it have anything to say about disequilibrium

R M WITCOMB

Cambridge

Incomes Policy and Inflation. Edited by MICHAEL PARKIN and MICHAEL T. SUMNER. Toronto University of Toronto Press, 1973. Pp 283 $10.00

At a time when much of the Western world is enmeshed in the morass of high unemployment and politically unacceptable rates of inflation, it is not surprising that governments have looked to "incomes policies" as a panacea for their economic ailments Economists have warned that economic theory predicts that incomes policies are likely to have potentially costly allocative consequences, however, so long as incomes policies hold out the hope of tempering inflation for even a short period of time, they are likely to remain in the politician's arsenal of economic weapons

This welcome volume of essays empirically examines the issue of whether or not the U K experiments with incomes policies have had a measurable aggregative effect on the inflationary process It is a useful quantitative counterpart to the qualitative study of incomes policies in Western Europe authored by Ullman and Flanagan

The volume begins with a survey of the econometric evidence on the domestic consequences of incomes policy by Parkin, Sumner, and Jones It continues with 12 interrelated empirical studies (eight of these have been previously published) and ends somewhat incongruously with one of Harry Johnson's omnipresent "Notes," which the editors characterize as a "bonus" The survey recognizes that before any effort can be made to assess the effects of incomes policies on inflation one must first answer the question, "What determines the rate of inflation?" (p 1) Although the authors acknowledge that there is no satisfactory answer to this issue since "all the studies cited exhibit defects at the statistical level and none of them deals convincingly with the problem of simultaneous equation bias" (p 18), they nevertheless conclude that "incomes policy apparently has little effect either on the wage determination process or on the average rate of wage inflation" (p 13) and that "there is no evidence that incomes policy has produced a consistent change in the price equation" (p 18) While such conclusions may be seized upon by political adversaries of incomes policies, my own conjecture is that a reasonably open-minded policy maker reading this volume is more likely to lessen his attachment to academic econometricians than to incomes policies

The core of the volume is an oft-cited paper by Lipsey and Parkin (L-P) which was previously published in *Economica* and is entitled "Incomes Policy A Reappraisal" With the exception of the Breckling and Smith papers which predate the L-P study, all of the remaining empirical essays are essentially variations and reappraisals of the L-P "reappraisal," each focusing on a different aspect of the problem of evaluating the effects of incomes policy As the introductory essay candidly informs the reader, "Most studies of the price equation have wasted little time on theoretical niceties before getting down to the serious business of estimation." The L-P paper lays out a useful analytic framework for indicating how incomes-policy effects might be detected and proceeds to estimate price and pressures to reduce costs outweigh the costs of duplicated facilities in "natural monopoly" situations. The best theoretical innovation (although there is little competition in this area) is Smallwood's model of the automobile insurance industry Unfortunately, Smallwood's empirical work is not well designed to test the relevance of the model, and the remainder of his paper does not reflect the careful thinking apparent in other chapters of the book Certainly the best chapter in the volume is Eads's study of service competition among domestic airlines This paper is recommended not only by the quality of its contents, but also by the author's clarity of exposition.

Any collection of articles is potentially more valuable and informative than the sum of its parts Insights which are difficult to gain from reading a single article are often more easily gleaned by comparing and contrasting the styles, methodology, and conclusions of several authors This is potentially the case with the Phillips collection. An overview of the articles reveals a number of obvious similarities and differences The chapters of the book which collate the materials are missing, however, so the reader must be diligent or he will overlook the most valuable lessons of the book

A most striking similarity in methodology is apparent among the nine published studies Each author follows an identical and very simple conceptual framework—one which I believe is useful not only in assessing the impact of changes in regulation but for changes in public policy in general In an effort to arrive at meaningful policy conclusions, each author seeks to identify industry performance in each of three alternative states of the world the current regulated state, a state with some form of deregulation, and an "ideal" state If deregulation is perceived to move performance toward rather than away from the "ideal," then that form of deregulation is offered as a desirable policy Note that some notion of ideal performance, even if unattainable, is necessary in order to provide a benchmark for the relevant comparison between regulated and deregulated performance Furthermore, while the selection of an ideal level of performance is necessarily a normative task, each author accepts, at least implicitly, the same ideal—the Pareto criterion of economic efficiency

The most difficult and crucial step in implementing this simple conceptual experiment is the identification of industry performance in the counterfactual world of less stringent regulation Geographical differences in the stringency of regulation or quantifiable changes in the degree of regulation over time sometimes provide the opportunity to estimate directly certain dimensions of the likely performance of an industry under increased competition In the absence of such data, as is very often the case, the investigator must rely more heavily upon educated conjecture The basis of such conjecture which predominates the papers in this volume is the presumed relationship between market structure and industry performance Empirical inquiry is directed toward determining the level of scale economies and the probable level of market concentration with deregulated pricing and entry The belief that concentration and performance are somehow related is certainly traditional, but this tradition is currently under attack by respected economists who see no bases for the belief, either in theory or in fact Eads, for example, implies that an airline which had a "natural monopoly" in a certain city pair market would not be able to exert its monopoly power in a regime of open entry because of the constant threat of potential competitors It seems clear that the relationship between concentration and performance is not a simple one, and the authors would have been well advised to consider more explicitly the conditions of entry, especially the mobility and specificity of the required capital

Even more interesting than the similarities in methodology are the differences in recommendations and policy prescriptions concluded by the various authors. Perhaps it is not surprising that the strongest appeals for deregulation arise from the scholars who investigated regulatory performance in the areas of transportation and banking. Weiss and Primeaux are much more skeptical of the role of greater market competition in the field of electricity. Waverman and Mann applaud the efficacy and viability of recent procompetitive policies adopted by the FCC and SEC, respectively, but neither author avows that greater strides in promoting competition in these industries would be similarly appropriate

The potentially most interesting chapter of this volume is notable by its absence. That chapter would have raised some of the following questions How can a group of scholars, skillfully striving to answer the same question and using similar methodology, arrive at such disparate conclusions? Why does regulation seem to work so much better for electric companies than for airlines? Why have the SEC and FCC recently adopted policies in the interest of consumers while the ICC remains a strong protector of its industry, even in the face of a growing variety of critics and social losses estimated to be a minimum of $4 billion per year? Is there a single "economic theory of regulation" which can encompass the wide ranges of often contradictory regulatory activity across all industries?

One reason that the pieces of this puzzle are so difficult to fit together is that we have been misled We have been misled into believing that public policy reflects a "goal" and that this goal is something like the "public good," or, even more absurdly, economic efficiency Almarin Phillips misleads us in his introduction when he says, "[T]he economics of regulation is the economics of 'market failure', that is, it is concerned with failures of the market mechanism to allocate resources properly and with the regulatory devices, including the use of market forces, that may be employed to correct for these failures " I infer from this that the "goal" of regulation is economic efficiency If the reader accepts this goal, he will never unravel the inconsistencies of regulatory performance.

Economic efficiency is the *result*, not the goal, of the competitive interaction of self-interested individuals in economic markets When this interaction does not result in economic efficiency, the market is sometimes said to "fail " Similarly, public policy is the *result*, not the *goal*, of the interaction of these same individuals in political markets As evidenced by the authors in this volume, political markets for regulation sometimes fail They not only fail to achieve a rosy ideal such as economic efficiency, but they also fail to do as well as alternative economic markets In other instances, however, regulatory performance is observed to be much more satisfactory The economics of regulation must be, in large part, the study of the causes of the success and failure of political markets

While economists have made great advances in understanding economic markets, we still know little about how political markets allocate resources. Political markets constrain individual interaction differently than do economic markets A most obvious difference is that political exchanges need not be voluntary This may have important ramifications upon the incentives to participate in political markets, since the farmer who shuns the market on Sunday may find his crops have been sold just the same, and at a very low price indeed It is not clear how the machinations of political markets are resolved into an equilibrium set of policies, or even that such an equilibrium exists. It is clear that a better understanding of the political marketplace is necessary if we are ever to make sense of the apparent schizophrenia of public policy, a condition of which the disparate findings of these authorities on regulation are symptomatic

MICHAEL M MURPHY

Southern Methodist University

Small Open Economies. Their Structure and Policy Environment. By MARTIN F. J. PRACHOWNY.
Lexington, Mass , Toronto, and London. D C. Heath & Co , 1975

This short and modest book is a useful one, not least because it sets a standard for measuring the contribution of future dissertations on the general equilibrium analysis of trade-and-monetary-assets models using "the small-country assumption" and concentrating on variations in the price ratio of traded to nontraded goods

The first three chapters justify the small open economy model against P. J Lloyd's argument for specific-country analysis, analyze the factor and commodity markets, using the increasingly popular elegant algebra developed by R W. Jones, and complete the model with the asset markets, including the Argy-Porter extension of Mundell's analysis to incorporate the forward market and the elasticity of expectations, and a rather rudimentary treatment of interest on public debt (In general, the argument skims over some of the subtleties of government debt and of money as wealth)

The remaining three chapters deal with the standard problem of monetary and fiscal policies for stabilization, the most important novel conclusion to emerge being that, with nontraded goods and floating rates, the Phillips curve trade-off between temporary inflation and unemployment is less serious for a small open economy than for a large or closed economy, the problem of devaluation, and the problem of inflation coming from abroad There is a significant error or misleading statement on page 111, where three cases are given when exchange rate policy can be useful External inflation is a valid case, but extra international reserves could be acquired equally well by domestic credit restraint, and forcing domestic residents to save more and reduce foreign holdings of government debt could be achieved in many easier ways than by devaluation (this way involves cheating the foreigner, a point not mentioned) In fact, in this kind of model, as is well known, devaluation and domestic credit contraction are perfect substitutes in controlling the level of reserves, exchange rate and monetary policy can differ only in dealing with external shocks

HARRY G JOHNSON

University of Chicago

Irish Land and British Politics Tenant-Right and Nationality, 1865–1870 By E D STEELE
Cambridge Cambridge University Press, 1974 Pp. ix-367 $19.50

When an Ulster farmer set out to rent a farm in the nineteenth century, he did not approach the landlord with an offer When a landlord wanted to let a holding, he did not auction it off to the highest competitive bidder Between the landlord and the prospective tenant stood the occupier It was to him that the prospective tenant made an offer If the holding was worth a certain amount to this prospective tenant, he would subtract the rent due the landlord and offer the rest to the occupier If agreeable, the occupier would pocket the sum and leave This sum was his tenant right, and it was guaranteed him by the custom of the country. The custom existed in various forms all over Ireland (though large tenants of prosperous farms typically negotiated leases directly with landlords).

It is obvious that the value of the tenant right depends on the level of rent and that a competitive rent, arrived at by auction to the highest bidder, would reduce it to zero. Tenant right depends on landlord acceptance of less-than-competitive rent. It is clearly in the pecuniary interest of a prospective tenant to offer such a landlord a rent in excess of that being paid by the occupier, and in the landlord's interest to accept it. What prevented such outcomes? The answer is that threats of violence from the organized occupying tenantry prevented landlords and "land grabbers" from reaching such accommodation. Threats and shots enforced the custom.

When Gladstone determined to legislate on Irish land in his first ministry, he concluded that the best procedure was to legalize this custom where it existed and to extend it where it did not. Since in the latter case no prospective tenants were in the habit of bidding on tenant right, the solution was for courts to determine what a hypothetical offer might be and to award that to the occupier. However, Gladstone failed to persuade his cabinet of the hypothetical offer idea, and when the bill emerged it contained instead a scale of payments to evicted tenants as compensation for disturbance (for reasons other than nonpayment of rent). The bill, which contained some other minor clauses, passed without difficulty.

In *Irish Land and British Politics*, Dr. Steele presents a summary of Gladstone's memoranda and correspondence on the bill, the memoranda and correspondence of members of his cabinet, the gist of contemporary newspaper and periodical comment, and a selection of opinion from other politicians' speeches and writings, including that of the Tory opposition. The Watergate tapes are a trickle to this flood of Victorian prose. The prose is better, the intellectual level higher, and the best lines and the best minds are the politicians', while the newspapermen put on a poor show. It shakes one's faith in progress.

What is to be learned from plowing through all these papers with Steele? He does not present his work as an essay in the "high politics" school of Maurice Cowling and others, who view jockeying for place and power as the "real" motivation behind legislation. Instead Steele takes at its face value much of the argumentation he reports. Does he view as significant findings the new light he casts on the role of Sullivan in drafting the bill? Of Lady Waldegrave in nagging her husband? Of Walters and Delane in shaping *Times* policy? Steele gives us very little guidance about what he sees as significant in all this mass of papers. He presents his conclusions in three pages. (1) The Land Act of 1870 was a total flop. (2) It "did not pacify Ireland" and in parts "it was a monumental irrelevance" (pp. 312-13). (3) Yet for England the results were astounding and revolutionary: "it was no accident . . . that the trickle of collectivist legislation passed to cope with the pressures of a growing industrial and urban society started to broaden into a stream from the year 1870," for the Irish Land Act of 1870 had sacrificed "the notion that the individual's rights of property were indefeasible, and with that cherished notion the concomitant freedom of contract" (p. 315).

Nothing in the previous 311 pages prepares us for these conclusions. What happened? Why did the Act fail? Did Gladstone flunk Decision Making I? The third conclusion I find unacceptable as a proposition in legal history and as a proposition about interventionist legislation; in any case, no piece of evidence is adduced in this book to support these astonishing claims.

What did Gladstone think was wrong in Ireland, and how did he think tenant right would fix it? He thought the Irish tenant wanted a share in his holding, which he felt the English had taken by conquest, and that the failure of law to define and protect the tenant's interest was the cause of Irish disaffection. This appears to be Steele's view. Nationality and tenant right are thus fused into one undifferentiated Irish protest. In fact, in mid-century those people who wanted national independence for Ireland had notoriously very little interest in land reform, and the tenants had notoriously very little interest in nationalist politics. No abstract legal system of land tenure commanded tenant support for its own sake: the tenants wanted to keep rents low. The economic problem of Ireland was less who owned what, it was poverty. It took considerable political skill to form a coalition of nationalist and peasant in Ireland. The architects of this coalition understood the distinctness of the nationalist and tenant movements as Gladstone did not. That's why they won.

There is ample evidence that Gladstone never understood how tenant right worked, what it was a payment for, and how its amount was determined. He confessed this often (pp. 121, 124, 126, 142, 250). Implacably against rent control, he did not see that fixity of tenure or free sale were meaningless if rent varied. The criticisms and arguments of his cabinet colleagues, many of which were logically valid and politically prescient, did not deter him. For example, John Bright, who had an accurate analysis of Irish politics and prescribed the very land policy that eventually was successfully adopted (by the Tories), was simply told in Gladstonian language to shut up.

In writing about Irish matters mainly from the papers of English statesmen, Steele stands in a time-honored tradition. To English statesmen the Irish were a troublesome race, prone to violence, continuously seeking independence under a parade of leaders from Brian Boru to Emmet and Tone, to O'Connell and Parnell, to Pearse and MacDonagh (notwithstanding a few differences of aims, methods, perceptions, class, and religion among them). That Ireland passed through a demographic, agrarian, social, and political metamorphosis in the nineteenth century that deserves to be called revolutionary was not perceived in the papers of English statesmen. No narrative that fails to discuss these distinctions and these changes can hope to tell us where Gladstone went wrong and why the Land Act of 1870 failed.

BARBARA L. SOLOW

Boston University

Income Distribution: Analysis and Policies. By JAN TINBERGEN.
Amsterdam: North Holland Publishing Co., 1975. Pp. 170. Fl 40.00 (about US$16.75).

Jan Tinbergen has been a leading figure in the development of economics and econometrics for several decades. Recently he also has been actively involved in Dutch social reform politics. His most recent book, *Income Distribution: Analysis and Policies,* can be expected to have an important place in the income distribution literature. The book surveys some of the economics literature, presents original research, and discusses public policies concerning the determinants of the distribution of labor market income (earnings) in developed economies. The analysis is highly original and creative. While the specifics of the treatment are sometimes troublesome, the book is always stimulating.

The book surveys the literature on trends in the inequality of individual and family income in developed countries. The general pattern that emerges is one of a long-term decline in inequality, including the post–World War II period, although the data end at 1966. This finding, however, warrants more words of caution than are offered. The data prior to the 1930s are few in number and of questionable quality and comparability to the more recent data. At least in part, the larger inequality during the Great Depression than in the post–World War II period may reflect a cyclical rather than a secular phenomenon. For the United

States, moreover, numerous studies indicating that there has been no secular trend in inequality in the last quarter-century are not mentioned in the survey. (See, for example, Chiswick and Mincer [1972], Metcalf [1969], and the references therein)

The theoretical analysis begins with a survey of the literature on two major approaches to income distribution analysis The "human capital school" has focused on the supply side of the labor market and generally views the wage structure as exogenous The "education planning" or "production function school" has focused on demand for different types of labor and derives relative marginal products from aggregate production functions, generally assuming that labor supplies are given Tinbergen notes correctly that the two approaches have seldom been combined to form a complete model of the distribution of labor income The book is devoted to a development of such a supply-and-demand model and its application to income inequality issues, including the effects of policy instruments

The model emphasizes the roles of human capital and economic development for explaining inequality The aggregate supply (or quantity) of human capital is represented by the average level of schooling or the proportion of the labor force with higher education The aggregate demand for human capital is postulated to be greater, the higher the level of income or the greater the proportion of the labor force in manufacturing The higher the level of schooling relative to the level of income, the lower would be the price of skill (relative wages) and the smaller the inequality of income This hypothesis is generally supported by cross-sectional regressions

Tinbergen believes that technological development tends to increase the relative wages of the educated and therefore income inequality "*unless a deliberate effort* be made to *direct technological innovation* to more labour-intensive methods in the sense of their requiring more *labour of low education level*" (p 117, italics in original) He therefore views the observed time series and cross-sectional pattern for income inequality as the outcome of a "race" between technology (income, development) and education—a race that education appears to have been "winning " Yet, this view requires assumptions about the skill neutrality of technical change to which his tests do not directly relate Nor does he inquire into the efficiency implications of his policy recommendation

The analysis of the relative supply of high-level manpower uses a utility function A household's utility (W_t) is postulated to be a function of its level of income (X_t), the education level "required" for the job held by the household head (S_t), and the squared difference between the required schooling level and the actual schooling level in the job of the household head [$(S_t - V_t)^2$], where $W_t = \ln(X_t + C_0S_t + C_1(S_t - V_t)^2)$ After assuming that all households have the same utility, ΔX is effectively regressed on ΔS and $\Delta(S - V)^2$ for grouped data to estimate the coefficients of the utility equation Note that the utility function, which is viewed as a supply relation, implies decreasing marginal utility of income It is not surprising therefore that optimal income inequality in the model implies less inequality than currently exists

Tinbergen is candid, however, concerning his views on utility functions "Despite all inequalities a hard core of 'fundamental equality' remains as has always been felt by religious and political idealists A more down-to-earth interpretation of this 'fundamental equality' or 'equivalence' is also conceivable the fact that each human being is a member of the human race in contradistinction to other living beings What is added in this study is a concrete *mathematical* expression of 'fundamental equality' or 'equivalence' by postulating that the utility functions of all human beings have the same mathematical shape and the same coefficients Our positive suggestion consists of adhering to the definition that equity stands for *equal welfare for all individuals*" (pp. 129–30, italics in original). Thus, he combines his normative and positive economics.

This reviewer, and most economists, believe that at most ordinal, but not cardinal, measures of utility could, in principle, be generated for households, and that interhousehold utility comparisons are not meaningful. Tinbergen is aware that his utility analysis will not be well received: "No doubt the greatest controversy about this book will concentrate on this issue. It would not be difficult, even for this author, to write a convincing and killing critique and many such reviews shall be written" (p. 150).

The discussion of the relative demand for labor inputs distinguishes "between the schooling normally required for the execution of a given productive task and the *actual* schooling of the person engaged for that task" (p. 80, italics in original). The input into the aggregate production function for schooling level 1 is not the number of persons with that schooling level (Φ_{1}) but $\Phi_{11} + \Pi_{21}\Phi_{21}$, where Φ_{ij} is the number of persons in job i with schooling level j ($i = 1,2$ and $j = 1,2$), Π_{ij} is the productivity ratio between persons in job i with schooling j (Π_{11} is by definition equal to unity) and $\Phi_{1} = \Phi_{11} + \Phi_{21}$. The rationale for this innovative attempt to bring quality dimensions into the production function is that "the productivity of an individual with an education level not intended for the job will differ from the productivity of an individual with an appropriate level of education for the job considered" (p. 82).

One problem with this approach is that the magnitude of Π_{21} and the division of Φ_{1} into Φ_{11} and Φ_{21} are both endogenous to the production process. In addition, if all persons in education 1 are perfect substitutes for each other and workers are paid their marginal products, competition would force $\Pi_{21} = 1$. If Φ_{11} and Φ_{21} persons are not perfect substitutes on a one-for-one basis because of barriers to job mobility, including a different quality of training, they could be viewed as separate factors. Or an explicit measure of the quality difference could be included in the analysis. These issues are not raised in the book.

An important parameter in a supply-and-demand model for income inequality is the elasticity of substitution between different types of manpower. The effect of an increase in the ratio of high level to other manpower on relative wages is greater, the smaller the elasticity of substitution between different skill levels in the aggregate production function. Tinbergen discusses cross-sectional estimates by himself and others of this elasticity. Most estimates have been obtained by assuming that the relative wage of skilled manpower is the dependent variable in the demand function and that relative quantity is an exogenous variable in the short run. Tinbergen argues correctly that this is appropriate for estimating a short-run elasticity of substitution but not necessarily for a long-run analysis. He believes that for long-run decisions "the organizers of production will start from their information on prices of products as well as production factors, including those of various types of labour" (p. 92), and therefore relative quantity should be the dependent variable. This conclusion is based on the implicit assumption that the long-run aggregate production function is the sum of individual firm production functions. Although the action of each firm separately is insignificant, the combined behavior of all firms affects relative wages. Analytically, in the long run both relative price and relative quantity are endogenous variables and the elasticity of substitution must be estimated using simultaneous equation techniques.

The empirical estimation of the elasticity of substitution between high level and other manpower is still an open issue. Tinbergen estimates the elasticity across the states of the United States using indirect least squares. He uses median years of schooling in the state as the supply shift variable and the proportion of the labor force in white-collar service occupations and transportation as the demand shift variables. It would seem, however, that these variables should be endogenous, rather than exogenous, in a model designed to explain the ratio of high level (defined by schooling level) to all manpower. His computed elasticity of slightly less than unity may be a consequence of this simultaneous equations bias toward zero. The Ullman result that Tinbergen compares with his own cross-state estimate is based on a two-stage procedure using cleaner exogenous variables: Ullman (1972) finds her data consistent with a CES production function, with an elasticity of substitution of 2.5 between high level and other manpower. The higher elasticity implies a much smaller narrowing of wage ratios for the same policy of increasing the proportion of college graduates in the labor force.

The book does not share the concern of some that "too many" college-trained persons have entered the labor force (pp. 104–5). Because he believes in price-responsive demand-and-supply curves and that structural unemployment does not exist in competitive labor markets, Tinbergen sees no problem. However, the analysis should be pushed one more step. If the increased current and future supply of college graduates is a consequence of public subsidies for schooling, as he implies it should be, there may be no inefficiency in private decision making, yet substantial *social* inefficiency in the allocation of the nation's resources. The social cost of schooling is not explicitly included in the model, and the implications of the policy recommendations for economic efficiency are again neglected.

In summary, Tinbergen's book contains many interesting, innovative, and stimulating ideas. However, the review of the literature suffers from being too selective on some important points, and some key aspects of the theoretical analysis and empirical interpretations are questionable. The quality of the presentation seems to suffer unduly from the reformer's zeal. And, the reader's task is made more difficult by editorial problems: the often awkward English, the frequent change in symbols from chapter to chapter for the same concept, and the absence of *t*-ratios, standard errors, or other indicators of statistical significance for the coefficients in many single-equation multiple regressions. Yet, for all of its faults, the book is to be welcomed as a thoughtful contribution on an important and perplexing set of problems.

BARRY R. CHISWICK

Council of Economic Advisers

References

Chiswick, Barry R., and Mincer, Jacob. "Time Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985." *J.P.E.* 80, no. 3, suppl. (May/June 1972): S34–S66.

Metcalf, Charles. "The Size Distribution of Personal Income during the Business Cycle." *A.E.R.* (September 1969), pp. 657–68.

Ullman, Carmel J. "The Growth of Professional Occupations in the American Labor Force: 1900 to 1963." Paper based on Columbia Univ. Ph.D. dissertation (1972), available from the World Bank.

JOURNAL OF MONEY, CREDIT AND BANKING

William G Dewald, EDITOR

ADVISORY BOARD

Martin Bailey
Phillip Cagan
James Duesenberry
Edward J Kane
Harry Markowitz
Franco Modigliani
Jurg Niehans
James Tobin

ASSOCIATE EDITORS

Ernst Baltensperger
William Brainard
Michael J Hamburger
H Robert Heller
Donald D Hester
Benjamin J Klebaner
Arie Melnik
Robert C Merton
Charles Nelson
William Poole
Robert A Rasche

Founded in 1969, the JMCB serves as a leading scientific journal in its field It has provided a medium of growing importance for presentation of research findings and policy evaluation in the economics of both domestic and international money and financial institutions Some articles are highly technical and others are designed to be understood by the general professional reader Contributors are drawn not only from North America, but from throughout the world

Topic areas include monetary and fiscal policy, regulation of financial institutions, money and credit markets, portfolio management, international money and balance of payments, financial aspects of economic development, and monetary history

In addition to full-length articles, the Journal features comments and replies reviews of selected books, and essay-length analyses of governmental and institutional reports Special issues are published occasionally and are included in annual subscriptions

Journal of Money, Credit, and Banking
Published quarterly in February, May, August, and November **Subscriptions** one year $16 00 (students, $10 00), additional subscriptions in the same name, $10 00 Two years $29 00 Three years $42 00 Outside the U S add $1 25 postage Single copies $4 50, back issues available Orders should be sent to the Ohio State University Press, 2070 Neil Avenue, Columbus, Ohio 43210 **Manuscripts** should be submitted in triplicate to William G Dewald, Editor, Department of Economics, Ohio State University, Columbus, Ohio 43210

Inflation Uncertainty and Expected Returns on Treasury Bills

Eugene F. Fama

University of Chicago

At time $t-1$ there is uncertainty about the rate of inflation that will be observed between $t-1$ and t, and there is uncertainty about the expected values of future rates of inflation that will be assessed at t The attempts to identify variation through time in expected real returns on Treasury bills as a function of the uncertainty of the inflation rate are unsuccessful It is argued, however, that uncertainty of the inflation rate doesn't change much during the sample period (1953–71) Variation through time in uncertainty about future assessments of expected rates of inflation seems somewhat more systematic, and perhaps as a consequence relationships between this source of inflation uncertainty and expected real returns and premiums are found

I. Introduction

Fama (1975) concludes that the behavior of returns on U S Treasury bills is consistent with the hypotheses that (*i*) expected real returns on bills are constant through time, and (*ii*) in setting the prices of bills the market is efficient in the sense that any information on past rates of change in the purchasing power of money is correctly used to assess expected future rates of change These conclusions are based on the findings that autocorrelations of real returns on bills are close to zero, and regressions of rates of change in the purchasing power of money on nominal interest rates indicate that all variation through time in nominal rates mirrors variation in correctly assessed expected values of future rates of change in purchasing power

This paper reexamines the hypothesis that expected real returns on bills are constant through time We test for relationships between expected

This work was supported by the National Science Foundation Donald Garren provided both intellectual and computational assistance, and I have had helpful comments from Nicholas Gonedes, Reuben Kessel, Richard Roll, and a referee

[*Journal of Political Economy*, 1976, vol 84, no 3]
© 1976 by The University of Chicago All rights reserved

real returns and the uncertainty in real returns that can be traced to uncertainty about inflation. The proposition is that the inflation-related uncertainties in the real returns on bills are associated with the type of risk that the market compensates.

This hypothesis is consistent with the early term structure literature, for example, Hicks (1946) and Kessel (1965), where the risks of bonds are equated with the uncertainty in their returns. In modern Markowitzian (1959) portfolio theory, however, risk and uncertainty of return are only equivalent for portfolios. The risk of a security in a portfolio is in part determined by the uncertainty of the security's return, but its contribution to the uncertainty in the return on a diversified portfolio usually depends more on how its return covaries with the returns on other securities.

The tests presented here nevertheless turn out to be interesting from the viewpoint of portfolio theory. It is, however, better to leave justification of this claim to the end of the paper where, with the evidence in hand, the issues are better appreciated.

II. Returns, Expected Returns, and Premiums

The first step is to present a system of notation for the bill pricing process which then helps us to formulate testable propositions about how the expected returns on bills might be related to different sources of inflation uncertainty. Since the empirical tests are based on monthly data, the model is stated in terms of monthly intervals.

The Pricing of Treasury Bills: Some Notation

Let $v_{n,t-1}$ be the price at the end of month $t-1$ of a bill with n months to maturity at (the end of the month) $t-1$. Consider a bill with one month to maturity at $t-1$, and assume that the maturity value of the bill at t, v_{0t}, is $1.00. The price of the bill at $t-1$ can be written as

$$v_{1,t-1} = e^{-R_t}, \tag{1}$$

where R_t is the sure, continuously compounded return on the bill from $t-1$ to t and is henceforth called the nominal or spot rate of interest for month t.

To describe the price of a bill that has 2 months to maturity at $t-1$, note that the bill becomes a 1-month bill at time t, so that

$$\tilde{v}_{1t} = e^{-\tilde{R}_{t+1}}. \tag{2}$$

The tildes indicate that $\tilde{v}_{1t}$ and $\tilde{R}_{t+1}$ are random variables at $t-1$. The expected value of $\tilde{H}_{2t}$, the return on the 2-month bill from $t-1$ to t, is then

$$E_{t-1}(\tilde{H}_{2t}) = E_{t-1}[\ln(\tilde{v}_{1t}/v_{2,t-1})] = -E_{t-1}(\tilde{R}_{t+1}) - \ln(v_{2,t-1}), \tag{3}$$

where $E_{t-1}(\tilde{R}_{t+1})$ is the expected value at $t - 1$ of the spot rate for month $t + 1$ to be set in the market at t. With (3), the price of a 2-month bill at $t - 1$ is

$$v_{2,t-1} = e^{-E_{t-1}(\tilde{H}_{2t}) - E_{t-1}(\tilde{R}_{t+1})}. \tag{4}$$

An expression for the price of a 3-month bill at $t - 1$ is obtained in similar fashion. Since the bill becomes a 2-month bill at t, its price at t can be expressed as

$$\tilde{v}_{2t} = e^{-\tilde{E}_t(\tilde{H}_{2,t+1}) - \tilde{E}_t(\tilde{R}_{t+2})} \tag{5}$$

The tildes indicate that quantities, including means of probability distributions, to be observed at t can be random variables at $t - 1$. From (5), the expected return from $t - 1$ to t on a 3-month bill is[1]

$$\begin{aligned} E_{t-1}(\tilde{H}_{3t}) &= E_{t-1}[\ln(\tilde{v}_{2t}/v_{3,t-1})] \\ &= -E_{t-1}(\tilde{H}_{2,t+1}) - E_{t-1}(\tilde{R}_{t+2}) - \ln(v_{3,t-1}), \end{aligned} \tag{6}$$

and the price of the bill at $t - 1$ is

$$v_{3,t-1} = e^{-E_{t-1}(\tilde{H}_{3t}) - E_{t-1}(\tilde{H}_{2,t+1}) - E_{t-1}(\tilde{R}_{t+2})} \tag{7}$$

The recursive nature of the arguments is clear For any $n \geq 1$, the price at $t - 1$ of a bill with n months to maturity is

$$v_{n,t-1} = e^{-E_{t-1}(\tilde{H}_{nt}) - E_{t-1}(\tilde{H}_{n-1,t+1}) - \qquad - E_{t-1}(\tilde{H}_{2,t+n-2}) - E_{t-1}(\tilde{R}_{t+n-1})} \tag{8}$$

Alternatively, we can introduce notation to recognize that $\tilde{H}_{k\tau}$, the return for month τ on a bill with $k > 1$ months to maturity at $\tau - 1$, is likely to differ from the spot rate R_τ Define the "premium" $\tilde{P}_{k\tau}$ as

$$\tilde{P}_{k\tau} = \tilde{H}_{k\tau} - R_\tau \tag{9}$$

Equation (8) can then be rewritten as

$$\begin{aligned} v_{n,t-1} = e^{-[E_{t-1}(\tilde{P}_{nt}) + R_t] - [E_{t-1}(\tilde{P}_{n-1,t+1}) + E_{t-1}(\tilde{R}_{t+1})] - } & \\ {}^{- [E_{t-1}(\tilde{P}_{2,t+n-2}) + E_{t-1}(\tilde{R}_{t+n-2})] - E_{t-1}(\tilde{R}_{t+n-1})} & \end{aligned} \tag{10}$$

Equation (8) expresses the current price of an n-month bill as the present value of the terminal \$1 00 to be received on the bill at $t + n - 1$, where the discount rate for any period is the current expected value of the return on the bill for that period Equation (10) breaks these expected returns into expected premiums and expected future spot rates

---

[1] Consistent assessment of probabilities implies that for any random variable $\tilde{x}_{t+\tau+k}$, $E_{t-1}[\tilde{E}_{t+k}(\tilde{x}_{t+\tau+k})] = E_{t-1}(\tilde{x}_{t+\tau+k})$, $\tau > 0$ and $k \geq 0$ See, e.g., Samuelson (1965)

The Uncertainties in Treasury Bill Returns

Assume that investors invest in order eventually to consume, so that they are concerned with real returns and prices There is a simple relationship between real and nominal returns If p_t is the money price at t of some "representative" bundle of consumption goods, then $\pi_t = 1/p_t$, the purchasing power of money at t, is the number of units of the bundle that can be purchased with one unit of money. The real return from $t - 1$ to t on a bill with n months to maturity at $t - 1$ is

$$\tilde{h}_{nt} = \ln (\tilde{\pi}_t \tilde{v}_{n-1,t}/\pi_{t-1} v_{n,t-1}) = \tilde{H}_{nt} + \tilde{\Delta}_t, \tag{11}$$

where $\tilde{\Delta}_t = \ln (\tilde{\pi}_t/\pi_{t-1})$ is the rate of change in purchasing power from $t - 1$ to t

When (11) is applied to a 1-month bill, we get

$$\tilde{r}_t \equiv \tilde{h}_{1t} = R_t + \tilde{\Delta}_t \tag{12}$$

Since the spot rate R_t is known at $t - 1$, the expected real return on a 1-month bill is

$$E_{t-1}(\tilde{r}_t) = R_t + E_{t-1}(\tilde{\Delta}_t) \tag{13}$$

Assume that equilibrium can be described in terms of expected real returns The pricing of a 1-month bill might now be described as follows At $t - 1$ the market assesses the distribution of $\tilde{\Delta}_t$, looks at other characteristics of the economic environment and then determines a market-clearing value of $E_{t-1}(\tilde{r}_t)$ The price $v_{1,t-1}$ and thus the spot rate R_t are then set so that the expected real return on a 1-month bill is equal to its equilibrium value, given the market's assessment of $E_{t-1}(\tilde{\Delta}_t)$ Since all uncertainty about $\tilde{r}_t$ is uncertainty about $\tilde{\Delta}_t$, if part of the value of $E_{t-1}(\tilde{r}_t)$ is a payment for bearing risk, then the risk must be associated with the uncertainty of $\tilde{\Delta}_t$

With (11) and (12), the real return and expected real return on an n-month bill are

$$\tilde{h}_{nt} = \tilde{r}_t + (\tilde{H}_{nt} - R_t) = \tilde{r}_t + \tilde{P}_{nt}, \tag{14}$$

and

$$E_{t-1}(\tilde{h}_{nt}) = E_{t-1}(\tilde{r}_t) + E_{t-1}(\tilde{H}_{nt} - R_t) = E_{t-1}(\tilde{r}_t) + E_{t-1}(\tilde{P}_{nt}) \tag{15}$$

With (15), the pricing of multimonth bills can be thought of as follows When the market sets the price of a 1-month bill, the first term $E_{t-1}(\tilde{r}_t)$ in the equilibrium expected real return on an n-month bill is determined, and the value of $E_{t-1}(\tilde{r}_t)$ includes any effects of uncertainty about $\tilde{\Delta}_t$ on $E_{t-1}(\tilde{h}_{nt})$ When the market sets the price of the n-month bill, the expected premium in $E_{t-1}(\tilde{h}_{nt})$ is determined If the expected premium contains compensation for the differential risk in the return on the n-month bill, then the differential risk must be associated with the uncertainty of the nominal return on the bill since, from (14), the uncertainty of $\tilde{h}_{nt}$ differs from the uncertainty of $\tilde{r}_t$ only because of the uncertainty of $\tilde{H}_{nt}$

Using (10) for the prices $\tilde{v}_{n-1,t}$ and $v_{n,t-1}$, $\tilde{H}_{nt}$ is

$$\begin{aligned}\tilde{H}_{nt} &= \ln (\tilde{v}_{n-1,t}/v_{n,t-1}) \\ &= E_{t-1}(\tilde{P}_{nt}) + R_t + [E_{t-1}(\tilde{P}_{n-1,t+1}) - \tilde{E}_t(\tilde{P}_{n-1,t+1})] \\ &\quad + [E_{t-1}(\tilde{R}_{t+1}) - \tilde{R}_{t+1}] + [E_{t-1}(\tilde{P}_{n-2,t+2}) - \tilde{E}_t(\tilde{P}_{n-2,t+2})] \\ &\quad + [E_{t-1}(\tilde{R}_{t+2}) - \tilde{E}_t(\tilde{R}_{t+2})] + \cdots + [E_{t-1}(\tilde{P}_{2,t+n-2}) \\ &\quad - \tilde{E}_t(\tilde{P}_{2,t+n-2})] + [E_{t-1}(\tilde{R}_{t+n-2}) - \tilde{E}_t(\tilde{R}_{t+n-2})] \\ &\quad + [E_{t-1}(\tilde{R}_{t+n-1}) - \tilde{E}_t(\tilde{R}_{t+n-1})]\end{aligned} \tag{16}$$

Since (13) holds for any t, it can be used to replace each of the future spot rates $\tilde{R}_{t+\tau}$ in (16) with a future expected real return $\tilde{E}_{t+\tau-1}(\tilde{r}_{t+\tau})$ and a future expected rate of change in purchasing power $\tilde{E}_{t+\tau-1}(\tilde{\Delta}_{t+\tau})$ Then, making use of the result in footnote 1, (16) becomes

$$\begin{aligned}\tilde{H}_{nt} - R_t = E_{t-1}(\tilde{P}_{nt}) &+ [E_{t-1}(\tilde{P}_{n-1,t+1}) - \tilde{E}_t(\tilde{P}_{n-1,t+1})] + \cdots \\ &+ [E_{t-1}(\tilde{P}_{2,t+n-2}) - \tilde{E}_t(\tilde{P}_{2,t+n-2})] \\ &+ [E_{t-1}(\tilde{r}_{t+1}) - \tilde{E}_t(\tilde{r}_{t+1})] + \cdots \\ &+ [E_{t-1}(\tilde{r}_{t+n-1}) - \tilde{E}_t(\tilde{r}_{t+n-1})] \\ &- [E_{t-1}(\tilde{\Delta}_{t+1}) - \tilde{E}_t(\tilde{\Delta}_{t+1})] - \cdots \\ &- [E_{t-1}(\tilde{\Delta}_{t+n-1}) - \tilde{E}_t(\tilde{\Delta}_{t+n-1})]\end{aligned} \tag{17}$$

Equation (17) identifies three possible sources of uncertainty in the premiums and thus in the nominal returns on multimonth bills There is uncertainty about the changes from $t - 1$ to t in (*i*) the expected values of the premiums in the future expected returns on the bill, (*ii*) the expected future real returns on 1-month bills, and (*iii*) the expected values of the rates of change in purchasing power in future months All three sources of uncertainty are possible candidates for explaining how any expected premium $E_{t-1}(\tilde{P}_{nt})$ in $E_{t-1}(\tilde{h}_{nt})$ might be related to the risk of the bill Since this paper is concerned with relationships between inflation uncertainties and expected returns, we concentrate on relationships between $E_{t-1}(\tilde{P}_{nt})$ and the uncertainty at $t - 1$ about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau = 1, 2, \ldots, t + n - 1$

The Testable Propositions

The uncertainty in the real return on a 1-month bill for month t is the uncertainty of $\tilde{\Delta}_t$ The value of $\tilde{\Delta}_t$ to be observed at t can be expressed as

$$\tilde{\Delta}_t = E_{t-1}(\tilde{\Delta}_t) + \tilde{\xi}_t \tag{18}$$

The uncertainty of $\tilde{\Delta}_t$ is the uncertainty of $\tilde{\xi}_t$ If this uncertainty, label it $\sigma(\tilde{\xi}_t)$, is associated with the type of risk that the market compensates, then part of the equilibrium expected real return on a 1-month bill is associated with $\sigma(\tilde{\xi}_t)$ Given some way to measure $\sigma(\tilde{\xi}_t)$ and variation through time in $\sigma(\tilde{\xi}_t)$, this is a testable proposition about $E_{t-1}(\tilde{r}_t)$

When the real return on an n-month bill and its expected value are expressed as in (14) and (15), $\tilde{\Delta}_t$ is included in $\tilde{r}_t$ and the effects of uncertainty about $\tilde{\Delta}_t$ on the equilibrium expected value of $\tilde{h}_{nt}$ are included in $E_{t-1}(\tilde{r}_t)$ To get additional propositions about relationships between expected real returns on multimonth bills and inflation uncertainties, we look to the premiums At $t - 1$ the market assesses $E_{t-1}(\tilde{\Delta}_t)$, and it also assesses expected values for future rates of change in purchasing power If these expected values change from $t - 1$ to t, then at $t - 1$, $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$, are random variables which, from (17), produce uncertainty in the premiums in the returns on multimonth bills If this uncertainty is associated with the type of risk that the market compensates, then uncertainty at $t - 1$ about expected rates of change in purchasing power to be assessed at t might be a source of expected premiums in the returns on multimonth bills Given some way to measure the uncertainty of $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$, this is a testable proposition about the behavior of expected premiums

The uncertainty at $t - 1$ about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$, is not a source of uncertainty in the real return on a 1-month bill, but it might have an indirect effect on $E_{t-1}(\tilde{r}_t)$ Because the real returns on 1-month bills are not affected by the uncertainty of future expected rates of change in purchasing power, an increase in this source of uncertainty might make 1-month bills more desirable relative to multimonth bills, and so their equilibrium expected returns might fall

Caution is, however, in order The goal of the paper is to measure relationships between expected real returns on bills and the inflation uncertainties in returns The task is of interest because, as the results will indicate, inflation uncertainties are by far the dominant source of uncertainty in the real returns on bills Consistent with the limited goal, we present a body of notation that suggests possible relationships between equilibrium expected returns and inflation uncertainties But we do not present a model of the pricing process specific enought to infer that the relationships are necessarily positive For example, from the viewpoint of portfolio theory, the model presented does not rule out a world where inflation uncertainties covary negatively with other sources of uncertainty in the returns on diversified portfolios so that they are negatively associated with the type of risk the market compensates

III. The Data

Description

Monthly prices of 1–6-month Treasury bills are from the quote sheets of Salomon Brothers The estimate of the nominal price v_{nt} is the average of the bid and asked prices on the last trading day of month t for the bill that matures closest to the end of month $t + n$ The reciprocal of the Bureau of Labor Statistics Consumer Price Index (CPI) is used to measure π_t, the purchasing power of money at the end of month t

Some details of the calculations should be mentioned. First, bill prices are sampled on the last trading day of each month, which is not always the last day of the month If left unadjusted, the nominal return H_{nt} and the rate of change in purchasing power Δ_t might cover slightly different periods To avoid any problems this might cause, H_{nt} and Δ_t are adjusted to an equivalent 30 4-day basis Second, prices quoted on bills are for delivery usually 2 days after the day of the quote In the adjustment process described above, the number of days covered by H_{nt} is computed from delivery date to delivery date or to maturity in the case of a 1-month bill Finally, H_{nt} is the return from $t - 1$ to t on a bill with n months to maturity at $t - 1$ Thus the bill used to measure H_{nt} changes from month to month

For 1-3-month bills, the tests cover the period January 1953 through July 1971 Data for 4-6-month bills are only available from March 1959 The starting date January 1953 coincides with an upgrading of the CPI The number of items in the CPI increases substantially and monthly sampling of major items becomes more general Since the tests require good measures of inflation uncertainty, attention is restricted to post-1953 data Data problems also explain the choice of July 1971 as the terminal date Beginning with August 1971, the Nixon administration engaged in a series of attempts to fix and unfix prices When prices are controlled, the true costs of goods to consumers are higher than quoted prices because of shortages, queues, and other distortions, and the large price changes reported when the controls are lifted overstate true price changes.

Average Returns and Premiums Some Perspective

Part A of table 1 shows summary statistics for the continuously compounded monthly real returns h_{nt} For different maturities n, the table shows $\bar{h}_n$ and $s(h_n)$, the mean and standard deviation of h_{nt}, the t-statistic $t(\bar{h}_n)$, the ratio of $\bar{h}_n$ to its standard error, the first three autocorrelations of h_{nt}, labeled $\hat{\rho}_1(h_n)$, $\hat{\rho}_2(h_n)$, and $\hat{\rho}_3(h_n)$, and T, the number of observations on h_{nt} during the period Part B of the table shows the corresponding sample statistics for the premiums $P_{nt} = H_{nt} - R_t$, while Part C shows summary results for H_{nt} The nominal return H_{1t} is the spot rate R_t, likewise, in Part A, $h_{1t} = r_t$ Because of the large autocorrelations of H_{nt}, t-statistics for $\bar{H}_n$ are meaningless, and they are not shown

During the sample periods, the average real returns on all bills are positive and large relative to their standard errors Thus there is convincing statistical evidence that expected real returns are positive. The average premiums are also always positive and many standard errors from zero, and the average premiums are an increasing function of maturity n

TABLE 1

SUMMARY STATISTICS

A. REAL RETURNS h_{nt}

Period and Bill (n)	$\bar{h}_n$	$s(h_n)$	$t(\bar{h}_n)$	$\hat{\rho}_1(h_n)$	$\hat{\rho}_2(h_n)$	$\hat{\rho}_3(h_n)$	T
Jan 1953–July 1971							
1	00074	00196	5 65	09	13	–.02	223
2	00102	00197	7 70	11	.15	– 02	223
3	00118	00202	8 73	13	.15	– 03	223
Mar 1959–July 1971							
1	00091	00169	6 54	05	05	– 07	149
2	00121	00172	8 58	06	07	– 09	149
3	00138	00179	9 39	10	10	– 08	149
4	00131	00192	8 36	13	10	– 08	149
5	00157	00199	9 61	.15	16	– 03	149
6	00162	00219	9 04	12	14	– 02	149

B PREMIUMS $P_{nt} = H_{nt} - R_t$

Period and Bill (n)	$\overline{H_n - R}$	$s(H_n - R)$	$t(\overline{H_n - R})$	$\hat{\rho}_1(H_n - R)$	$\hat{\rho}_2(H_n - R)$	$\hat{\rho}_3(H_n - R)$	T
Jan 1953–July 1971							
2	00027	00031	13 15	38	20	17	223
3	00044	00056	11 67	31	14	20	223
Mar 1959–July 1971							
2	00030	00034	10 96	40	22	19	149
3	00047	00058	10 05	33	.18	29	149
4	00041	00079	6 30	26	06	11	149
5	00066	00104	7 76	33	.16	28	149
6	00072	00129	6 77	25	10	21	149

C NOMINAL RETURNS H_{nt}

Period and Bill (n)	$\bar{H}_n$	$s(H_n)$	$\hat{\rho}_1(H_n)$	$\hat{\rho}_2(H_n)$	$\hat{\rho}_3(H_n)$	T
Jan 1953–July 1971						
1	00263	00129	97	96	94	223
2	00290	00138	93	91	89	223
3	00307	00151	85	81	81	223
Mar 1959–July 1971						
1	00319	00115	96	94	92	149
2	00349	00123	90	88	86	149
3	00367	00136	82	77	78	149
4	00360	00149	73	67	67	149
5	00385	00159	64	56	60	149
6	00391	00173	51	42	47	149

The average real returns and premiums are, however, small relative to what one observes for more risky securities like common stocks For example, the average monthly real return on a 1-month bill for the March 1959–July 1971 period is $\bar{r} =$ 00091 or about 09 percent (which translates into about 1 0 percent per year), and the standard deviation of the monthly real return is $s(r) = .00169$ or about .17 percent. For the same period the average premium in the return on a 6-month bill is $\overline{H_6 - R} = .00072$, or about .07 percent per month, and $s(H_6 - R) = .00129$, or about .13 percent per month. In contrast, when one looks at long periods of time, the average return on an equally weighted portfolio of all New York Stock Exchange stocks is typically close to 1.0 percent per month, and the standard deviation of the monthly return is about 5 percent per month.[2]

Finally, equation (14) expresses the real return $\tilde{h}_{nt}$ on an n-month bill as the real return $\tilde{r}_t$ on a 1-month bill plus the premium $\tilde{H}_{nt} - R_t$. The sample correlations between $\tilde{r}_t$ and $\tilde{H}_{nt} - R_t$, which we do not show, are close to zero, so that

$$\sigma^2(\tilde{h}_{nt}) \cong \sigma^2(\tilde{r}_t) + \sigma^2(\tilde{H}_{nt} - R_t)$$

It is then interesting to note that the standard deviations of $H_{nt} - R_t$ in table 1 are usually small relative to the standard deviations of r_t, variation of $\tilde{r}_t$ through time is the primary source of the variation in $\tilde{h}_{nt}$. Since, from (12), all the uncertainty of $\tilde{r}_t$ is uncertainty about $\tilde{\Delta}_t$, this inflation uncertainty is also the primary source of uncertainty in the real return on any bill. We argue later that the remaining uncertainty of $\tilde{h}_{nt}$, the uncertainty of $\tilde{H}_{nt} - R_t$, is largely the uncertainty at $t - 1$ of $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau = 1, 2, \ldots, t + n - 1$. Thus the eventual conclusion is that almost all the uncertainty in the real returns on bills is inflation related.

IV. Measuring Inflation-related Uncertainties: General Approach

The finding that average real returns on bills are positive but small is not new (see Fama [1975]), nor is the conclusion that there are systematic premiums in the average returns on multiperiod bills (see, for example, Kessel [1965] or McCulloch [1975]). The new results that we seek concern possible relationships between the inflation uncertainties in the real returns on bills and the levels of expected real returns and premiums. The task requires measures of the inflation uncertainties.

There are two sources of inflation uncertainty in the real returns on bills. First, there is uncertainty at $t - 1$ about $\tilde{\xi}_t$ in (18), the deviation of $\tilde{\Delta}_t$ from $E_{t-1}(\tilde{\Delta}_t)$. Second, there is uncertainty at $t - 1$ about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$, the expected rates of change in purchasing power to be assessed at t. To measure these two types of inflation uncertainty, we rely on the results of Fama (1975), who concludes that, at least for the 1953–71 period, and as a good first approximation, the expected real return on a 1-month Treasury bill is constant through time. In this situation, the deviation of $\tilde{r}_t$ from its constant expected value is the deviation of $\tilde{\Delta}_t$ from its expected value, and the time series of $\tilde{r}_t$ can be used to estimate the uncertainty of $\tilde{\xi}_t$.

---

[2] See, e.g., Fama and MacBeth (1974), table 4.

From (13) we can also see that with a constant expected real return the variation through time of the spot rate R_t mirrors the variation of the expected rate of change in purchasing power $E_{t-1}(\tilde{\Delta}_t)$. In this world, the uncertainty at $t - 1$ about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$ is the same as the uncertainty about $\tilde{E}_t(\tilde{R}_{t+\tau})$. We now see that the uncertainty at $t - 1$ about $\tilde{E}_t(\tilde{R}_{t+\tau})$ can be measured by studying the time series properties of R_t.

V. Expected Premiums and Uncertainty about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$

The Spot Rate and Uncertainty about Expected Future Inflation Rates

Fama (1975) finds that during the 1953–71 period, the spot rate R_t behaves much like a martingale. The fact that the autocorrelations of $H_{1t} = R_t$ in table 1 are close to one is consistent with this conclusion. Thus, to a good approximation, at any time $t - 1$ the spot rate $\tilde{R}_{t+1}$ for month $t + 1$ that will be observed at t is

$$\tilde{R}_{t+1} = R_t + \tilde{\eta}_t, \tag{19}$$

where $\tilde{\eta}_t$, the change in the spot rate from $t - 1$ to t, has expected value equal to zero. It follows that (*i*) the spot rate R_t is the expected value at $t - 1$ of all future spot rates, (*ii*) the change in the spot rate from $t - 1$ to t is the change from $t - 1$ to t in the expected values of all future spot rates, and (*iii*) with a constant expected real return, the change in the spot rate also mirrors the change from $t - 1$ to t in expected future rates of change in purchasing power. The uncertainty of $\tilde{\eta}_t$ is thus the uncertainty at $t - 1$ about the future expected rates of change in purchasing power to be assessed at t.

The analysis suggests a testable proposition to explain why the average premiums in table 1 increase with maturity. If R_t follows a martingale and $E_{t-1}(\tilde{r}_t) = E(\tilde{r})$, all the changes $E_{t-1}(\tilde{\Delta}_{t+\tau}) - \tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau = 1, \ldots, t + n - 1$, in (17) are equal to $\tilde{R}_{t+1} - R_t$. Thus the change in the spot rate from $t - 1$ to t pushes the premiums on all bills in the same direction. We can see from (17), however, that the premium $\tilde{H}_{nt} - R_t$ contains more of the changes $E_{t-1}(\tilde{\Delta}_{t+\tau}) - \tilde{E}_t(\tilde{\Delta}_{t+\tau})$ the longer the maturity of the bill. It follows that uncertainty about the change in the spot rate creates more uncertainty in a bill's premium the longer the maturity of the bill. If this source of uncertainty is positively associated with the types of risks that the market compensates, then expected premiums on bills will be an increasing function of maturity, and any changes through time in the degree of uncertainty about changes in spot rates will have a greater effect on a bill's expected premium the longer the maturity of the bill.

Formally, let $\sigma(\tilde{\eta}_t)$ be some measure of the uncertainty of $\tilde{R}_{t+1} - R_t$, and suppose the relationship between the expected premium on an n-month bill and $\sigma(\tilde{\eta}_t)$ is

$$E_{t-1}[\tilde{H}_{nt} - R_t \mid \sigma(\tilde{\eta}_t)] = \alpha_{0n} + \alpha_{1n}\sigma(\tilde{\eta}_t) \tag{20}$$

The conditional expected value $E[\tilde{H}_{nt} - R_t \mid \sigma(\tilde{\eta}_t)]$ is the regression function of $\tilde{H}_{nt} - R_t$ on $\sigma(\tilde{\eta}_t)$. Given time series of $\tilde{H}_{nt} - R_t$ and $\sigma(\tilde{\eta}_t)$, we can use regression techniques to estimate α_{0n} and α_{1n}. The proposition suggested by the analysis above is that α_{1n} is positive and increasing in n.

Measuring the Uncertainty of the Future Spot Rate

Since $\sigma(\tilde{\eta}_t)$ is some measure of the uncertainty at $t - 1$ about the change in the spot rate from $t - 1$ to t, we can interpret $\sigma(\tilde{\eta}_t)$ as some dispersion parameter and use observed changes in the spot rate to estimate the parameter. There are, however, conflicting considerations. To get meaningful estimates of α_{0n} and α_{1n} in (20), the estimates of $\sigma(\tilde{\eta}_t)$ must be sensitive to changes in $\sigma(\tilde{\eta}_t)$ through time. This suggests that we not use "too many" observations on the change in the spot rate to estimate $\sigma(\tilde{\eta}_t)$. If "too few" observations are used, however, the estimates of $\sigma(\tilde{\eta}_t)$ will be subject to substantial sampling error with the attendant problems that this causes in estimating the regression coefficients in (20).

One approach to this problem is to allow estimates of (20) to select the procedure for estimating $\sigma(\tilde{\eta}_t)$. First define the dispersion parameter as $\sigma(\tilde{\eta}_t) = E|\tilde{R}_{t+1} - R_t|$, the first absolute moment of the distribution of the change in the spot rate.[3] The general formula used to estimate $\sigma(\tilde{\eta}_t)$ is then

$$s(\eta_t) = s(R_{t+1} - R_t) = \frac{|R_{t-k} - R_{t-k-1}| + \cdots + |R_{t+q} - R_{t+q-1}|}{k + q + 1} \tag{21}$$

To estimate the regression coefficients α_{0n} and α_{1n} in (20), $s(\eta_t)$ is substituted for $\sigma(\tilde{\eta}_t)$ and ordinary least squares is applied to the time series of $H_{nt} - R_t$ and $s(\eta_t)$. To see which version of $s(\eta_t)$ from (21) is "best", the regression coefficients in (20) are reestimated for different choices of k and q in (21). In one set of tests, only past changes in the spot rate are used to compute $s(\eta_t)$, in another, only future changes in the spot rate are used, and finally, $s(\eta_t)$ is computed from both past and future changes in the spot rate. Without going into details, when the coefficient of determination in the regressions of $H_{nt} - R_t$ on $s(\eta_t)$ is used as the criterion, the best version of $s(\eta_t)$ seems to be obtained with $k = 12$ past changes and $q = 12$ future changes in the spot rate. This is the version of $s(\eta_t)$ used in the tests below.

---

[3] Experiments with $E[(\tilde{R}_{t+1} - R_t)^2]$ yield results similar to those reported below.

The Relationships between $E(\tilde{H}_{nt} - R_t)$ and $s(\eta_t)$

Table 2 summarizes the time series regressions of $H_{nt} - R_t$ on $s(\eta_t)$ For each bill n, the table shows a_{0n} and a_{1n}, the estimates of the regression coefficients α_{0n} and α_{1n} in (20), $t(a_{0n})$ and $t(a_{1n})$, the ratios of the coefficients to their standard errors; the coefficient of determination, adjusted for degrees of freedom, $s(e)$, the standard error of the residuals from the estimated regression function, the first three autocorrelations of the residuals, and T, the number of observations used to estimate the regression coefficients Since past and future values of the change in the spot rate are used to obtain $s(\eta_t)$, some of the data from the beginning and end of the period are lost for the purpose of estimating the regression coefficients

Table 2 supports the hypotheses about the relationships between expected premiums and uncertainty about future expected rates of change in purchasing power The estimates a_{1n} of α_{1n} in (20) are positive for bills of all maturities, and generally more than three standard errors above zero Thus an increase (reduction) in uncertainty about the change from $t - 1$ to t in expected future inflation rates is associated with an increase (reduction) in the expected premiums in the returns on multimonth bills Since the coefficients a_{1n} generally increase with maturity n, as predicted, this source of inflation uncertainty has a larger effect on expected returns the longer the maturity of the bill

Uncertainty at $t - 1$ about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$, seems to explain entirely the levels of average premiums The intercept a_{0n} is the estimate of the expected premium in the return on an n-month bill when there is no uncertainty about future expected rates of change in purchasing power The values of a_{0n} are negative in some cases and positive in others, and they are usually less than one standard error from zero

There is, however, good evidence that the estimated regression functions do not explain all the variation in expected premiums The first-order autocorrelations $\hat{\rho}_1(e)$ of the regression residuals are usually positive and fairly large They are smaller than the autocorrelations of the premiums in table 1, but the autocorrelations of the residuals in table 2 nevertheless indicate that there is variation through time in expected premiums which is not picked up by the regression function estimates Somewhat curiously, the regressions seem to explain the generally substantial higher order autocorrelations of premiums in table 1

There is, of course, the possibility that the first-order autocorrelations of the regression residuals would also disappear with better estimates of the variation through time of uncertainty about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$ The measure $s(\eta_t)$ is the end result of much experimentation with such measures, but nevertheless it may not adapt quickly enough to true changes in the uncertainty of future expected inflation rates

TABLE 2

SUMMARY STATISTICS FOR PREMIUM REGRESSION TESTS $H_{nt} - R_t = a_{0n} + a_{1n}s(\eta_t) + e_t$

Period and Bill (n)	a_{0n}	a_{1n}	$t(a_{0n})$	$t(a_{1n})$	Coeff of Det	$s(e)$	$\hat{\rho}_1(e)$	$\hat{\rho}_2(e)$	$\hat{\rho}_3(e)$	T
Jan 1953–July 1971										
2	− 00007	1 51	−1 22	6 37	17	00028	26	04	02	198
3	− 00015	2 58	−1 54	6 29	16	00049	15	− 07	07	198
Mar 1959–July 1971										
2	00002	1 12	0 29	3 93	11	00029	36	04	− 05	124
3	− 00002	1 99	−0 20	4 68	15	00044	18	− 10	02	124
4	00004	1 50	0 24	2 34	04	00066	19	− 12	− 15	124
5	00003	2 39	0 14	3 01	06	00082	25	− 05	01	124
6	− 00011	3 15	−0 44	3 15	07	00103	15	− 11	− 09	124

Finally, Kessel (1965) hypothesizes that expected premiums are related to the level of interest rates. To see if this proposition might explain the residual autocorrelations in table 2, multiple regressions of $H_{nt} - R_t$ on both $s(\eta_t)$ and R_t were tried The coefficients of $s(\eta_t)$ were almost identical to those in table 2 The coefficients of R_t were neither systematically positive nor negative, and their t-statistics were generally less than 1.0. As one might then expect, the residual autocorrelations were almost identical to those in table 2

VI. Expected Real Returns and Uncertainty at $t - 1$ about $E_t(\tilde{\Delta}_{t+\tau})$

The earlier discussion suggests that if expected premiums are positively related to $\sigma(\tilde{\eta}_t)$, there might be a negative relationship between $E_{t-1}(\tilde{r}_t)$, the expected real return on a 1-month bill, and uncertainty at $t - 1$ about $E_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$ Tests of this proposition are presented in table 3 Table 3 is also concerned with the relationships between the expected real returns on multimonth bills and $\sigma(\tilde{\eta}_t)$ Since we hypothesize a positive relationship between expected premiums and $\sigma(\tilde{\eta}_t)$ and a negative relationship between $\sigma(\tilde{\eta}_t)$ and $E_{t-1}(\tilde{r}_t)$, we do not have an unambiguous proposition about the relationships between $E_{t-1}(\tilde{h}_{nt})$ and $\sigma(\tilde{\eta}_t)$ We can, however, examine what the data have to say about the question Formally, if the regression function $E_{t-1}[\tilde{h}_{nt} \mid \sigma(\tilde{\eta}_t)]$ is

$$E_{t-1}[\tilde{h}_{nt} \mid \sigma(\tilde{\eta}_t)] = \alpha_{0n} + \alpha_{1n}\sigma(\tilde{\eta}_t), \tag{22}$$

we can substitute for $\sigma(\tilde{\eta}_t)$ the estimate $s(\tilde{\eta}_t)$ from (21) with $k = q = 12$, and then estimate the regression coefficients α_{0n} and α_{1n} These are the regression estimates summarized in table 3

The two estimates a_{11} of α_{11} are negative and fairly large relative to their standard errors Thus the results are consistent with the posited negative relationship between $E_{t-1}(\tilde{r}_t)$ and $\sigma(\tilde{\eta}_t)$ The negative relationships between the expected real returns on 1-month bills and $s(\eta_t)$ seem, moreover, sufficient to negate the positive relationships observed in table 2 between $s(\eta_t)$ and expected premiums The result is that there are no reliable relationships between the expected real returns on multimonth bills and uncertainty about future expected rates of change in purchasing power For $n \geq 1$, there seems to be some evidence of negative relationships between $E_{t-1}(\tilde{h}_{nt})$ and $s(\eta_t)$ for the 1959–71 period For the entire precontrols period 1953–71, however, there is no evidence in table 2 that expected real returns on 2- and 3-month bills have any relationship with $s(\eta_t)$, for these bills, the t-statistics for a_{1n} are both less than 1 0

VII. Expected Real Returns and Uncertainty about $\tilde{\Delta}_t$

Equation (11) says that part of the uncertainty in the real return on any bill is uncertainty about the rate of change in purchasing power $\tilde{\Delta}_t$

TABLE 3

EXPECTED REAL RETURNS AND UNCERTAINTY ABOUT FUTURE EXPECTED INFLATION RATES REGRESSION TESTS

$$h_{nt} = a_{0n} + a_{1n}s(\eta_t) + e_t$$

Period and Bill (n)	a_{0n}	a_{1n}	$t(a_{0n})$	$t(a_{1n})$	Coeff of Det	$s(e)$	$\hat{\rho}_1(e)$	$\hat{\rho}_2(e)$	$\hat{\rho}_3(e)$
Jan 1953–July 1971									
1	00138	−2 73	3 60	−1 70	01	00191	.06	.15	−.04
2	00131	−1 22	3 39	−0 76	− 00	00193	09	18	−.03
3	00123	−0 15	3 12	−0 09	− 01	00196	11	.18	− 04
Mar 1959–July 1971									
1	00197	−4 74	5 30	−3 05	06	00160	− 03	00	− 13
2	00199	−3 61	5 29	−2 30	03	00162	00	02	−.12
3	00195	−2 74	5 11	−1 72	02	00165	04	.02	− 12
4	00200	−3 24	5 00	−1 93	02	00173	08	01	−.14
5	00199	−2 35	4 93	−1 39	01	00175	09	06	−.12
6	00186	−1 59	4 17	−0 85	− 00	00193	07	04	−.15

When, as in (14), $\tilde{h}_{nt}$ is expressed as the sum of $\tilde{r}_t$ and $\tilde{H}_{nt} - R_t$, the uncertainty about $\tilde{\Delta}_t$ is included in $\tilde{r}_t$ Thus if the uncertainty of $\tilde{\Delta}_t$ affects expected real returns, the effects should be contained in the expected real return on a 1-month bill We do not expect to observe relationships between expected premiums and uncertainty about $\tilde{\Delta}_t$

As discussed earlier, under certain conditions the uncertainty of $\tilde{\Delta}_t$, that is, the uncertainty of $\tilde{\xi}_t$ in (18), is the uncertainty of $\tilde{r}_t$ Define $\sigma(\tilde{\xi}_t) = E(|\tilde{\xi}_t|)$ The specific equation used to estimate $\sigma(\tilde{\xi}_t)$ is then

$$s(\xi_t) = \frac{|r_{t-12}| + \quad + |r_{t-1}| + |r_{t+1}| + \quad + |r_{t+12}|}{24} \tag{23}$$

If the regression function relationship between $E_{t-1}(\tilde{r}_t)$ and the uncertainty of $\tilde{\Delta}_t$ is

$$E_{t-1}[r_t \mid \sigma(\tilde{\xi}_t)] = \alpha_0 + \alpha_1 \sigma(\tilde{\xi}_t), \tag{24}$$

we can substitute $s(\xi_t)$ for $\sigma(\tilde{\xi}_t)$ and then estimate the regression coefficients α_0 and α_1 The estimated regression equation for the January 1953–July 1971 period is[4]

$$\begin{aligned} r_t = \; & \underset{(\ 45)}{00037} + \underset{(\ 49)}{233} s(\xi_t) + e_t, \end{aligned} \tag{25}$$

coeff of det	$s(e)$	$\hat{\rho}_1(e)$	$\hat{\rho}_2(e)$	$\hat{\rho}_3(e)$
− 00	00192	07	16	− 04

where t-statistics are shown in parentheses below the regression coefficient estimates, and other summary statistics are labeled as in tables 2 and 3 There is no evidence of any systematic relationship between $E_{t-1}(\tilde{r}_t)$ and $s(\xi_t)$ The estimate a_1 of α_1 is less than 5 standard errors from zero

Since, according to (17), uncertainty about $\tilde{\Delta}_t$ is not part of the uncertainty of the premium $\tilde{H}_{nt} - R_t$, we do not anticipate relationships between expected premiums and $s(\xi_t)$ To test this hypothesis, multiple regressions of premiums on both $s(\xi_t)$ and $s(\eta_t)$ were estimated The relationships between expected premiums and $s(\eta_t)$ observed in table 2 showed up, but the coefficients of $s(\xi_t)$ were generally less than one standard error from zero

VIII. Interpretation of the Results

The propositions about the relationships between $E_{t-1}(\tilde{H}_{nt} - R_t)$ or $E_{t-1}(\tilde{r}_t)$ and uncertainty at $t - 1$ about expected rates of change in purchasing power to be assessed at t are supported by the data, but the proposition that $E_{t-1}(\tilde{r}_t)$ is positively related to uncertainty at $t - 1$

---

[4] Experiments with different versions of (23), similar to those described earlier for (21), have been carried out The results in (25) are representative

TABLE 4

TIME SERIES BEHAVIOR OF INFLATION-RELATED UNCERTAINTIES, EXPECTED REAL RETURNS AND EXPECTED PREMIUMS

A. SKETCH OF THE TIME SERIES BEHAVIOR OF THE VARIABLES

Date	$s(\zeta_t)$	$s(\eta_t)$	f_t	$\widehat{H_{2t} - R_t}$	$\widehat{H_{3t} - R_t}$	R_t
Feb 1954	00192	00015	00097	00016	00023	0008
Feb 1955	00205	00016	00096	00017	00025	0009
Feb 1956	00206	00020	00083	00024	00037	.0019
Feb 1957	00187	00023	00077	00027	00043	0025
Feb 1958	.00173	00024	00074	00029	00046	0013
Feb 1959	00166	00031	00055	00039	00064	0020
Feb 1960	00178	00037	00036	00049	00081	0028
Aug 1960	.00185	00042*	00024†	00056*	00093*	0016
Feb 1961	00161	00035	00043	00045	00074	0017
Feb 1962	00174	00019	00086	00022	00034	0021
Feb 1963	00187	00015	00098	00015	00023	0024
Feb 1964	00185	00013	00104	00012	00017	0029
Mar 1965	00197	00011†	00107*	00010†	00014†	0033
Feb 1966	00189	00015	00099	00015	00022	0037
Feb 1967	00130	00018	00091	00019	00030	0037
Feb 1968	00105†	00024	00074	00029	00046	0040
Feb 1969	00134	00033	00049	00042	00069	0049
Feb 1970	00134	00031	00053	00040	00066	0063

* Maximum value
† Minimum value

B SUMMARY STATISTICS FOR LEVELS AND DIFFERENCES (d) OF VARIABLES

Statistic/x	$s(\zeta_t)$	$s(\eta_t)$	f_t	$\widehat{H_{2t} - R_t}$	$\widehat{H_{3t} - R_t}$	R_t
$\bar{x}$	001690	000236	000738	000287	000457	00280
$s(x)$	000307	000084	000228	000126	000216	00129
$\hat{\rho}_1(x)$	94	99	99	99	99	96
$\hat{\rho}_2(x)$	89	97	97	97	97	93
$\hat{\rho}_3(x)$	85	95	96	96	96	90
$\overline{dx}$	000020	000001	− 000003	000001	000002	000020
$s(dx)$	000325	000012	000033	000018	000032	000325
$\hat{\rho}_1(dx)$	− 23	14	14	14	14	− 23
$\hat{\rho}_2(dx)$	08	− 02	− 02	− 02	− 02	08
$\hat{\rho}_3(dx)$	02	10	10	10	10	02

about the rate of change in purchasing power from $t - 1$ to t gets no support To understand why the hypotheses about the two different sources of inflation uncertainty fare somewhat differently, we now examine the behavior through time of $s(\eta_t)$ and $s(\zeta_t)$ We also examine the importance of the measured variation through time of expected real returns and premiums

The Behavior through Time of $s(\eta_t)$ and $s(\zeta_t)$

Part A of table 4 shows $s(\eta_t)$ and $s(\zeta_t)$, generally at annual intervals There is much variation through time in $s(\eta_t)$, the measure of the uncertainty at $t - 1$ about $\hat{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$. The estimates start at .00015 in February 1954, almost triple in size to a high of .00042 in August 1960, fall to a low of .00011 in March 1965, and then rise back to about .00030 toward the end of the sampling period. In relative terms, there is much less variation in $s(\xi_t)$, the measure of the uncertainty at $t - 1$ about $\tilde{\Delta}_t$. In the 13 years from 1954 to 1966, $s(\xi_t)$ varies just slightly in the neighborhood of .00200. The estimates drop by about 30 percent in 1967 and stay at approximately the .00130 level until the end of the sampling period. One can argue, however, that even this apparent shift in $s(\xi_t)$ is a false indication of a change in uncertainty about $\tilde{\Delta}_t$. In 1965, there is an upgrading of the CPI. The geographical locations and goods for which prices are sampled increase. The consequent increase in the reliability of the CPI as a measure of the price level might account for much of the drop in $s(\xi_t)$ observed in following years.

The hypothesis that there is variation through time in expected real returns or premiums which is related to variation in the two sources of inflation uncertainty are only meaningful if there is variation in the inflation uncertainties. Thus, given the difference between the behavior of $s(\xi_t)$ and $s(\eta_t)$, it is not surprising that there is no evidence for hypotheses about the effects of variation in $\sigma(\tilde{\xi}_t)$ on $E_{t-1}(\tilde{r}_t)$, whereas there is evidence for the propositions about the effects of variation in $\sigma(\tilde{\eta}_t)$ on $E_{t-1}(\tilde{r}_t)$ and on $E_{t-1}(\tilde{H}_{nt} - R_t)$.

The Variation through Time of R_t and $E_{t-1}(\tilde{r}_t)$

The interpretation of $s(\eta_t)$ as a measure of uncertainty at $t - 1$ about $\hat{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$, is based on the assumption that the expected real return on a 1-month bill is constant through time, so that all variation in the spot rate R_t mirrors variation in $E_{t-1}(\tilde{\Delta}_t)$. The interpretation of $s(\xi_t)$ as a measure of uncertainty about $\tilde{\Delta}_t$ is likewise based on the assumption that $E_{t-1}(\tilde{r}_t)$ is constant so that all variation through time in $\tilde{r}_t$ reflects the deviations of $\tilde{\Delta}_t$ from $E_{t-1}(\tilde{\Delta}_t)$. The results for 1-month bills in table 3 indicate, however, that there is a measurable negative relationship between $E_{t-1}(\tilde{r}_t)$ and $s(\eta_t)$.

Changes through time in $E_{t-1}(\tilde{r}_t)$ do not invalidate the testing procedures unless the changes are large relative to the inflation uncertainties we wish to measure. Thus, $s(\xi_t)$ is a good measure of the uncertainty of $\tilde{\Delta}_t$ as long as variation in $E_{t-1}(\tilde{r}_t)$ is a small fraction of the variation in $\tilde{r}_t$. In table 3, the coefficient of determination for the regression of $\tilde{r}_t$ on $s(\eta_t)$ for the 1953–71 period is .01. Since only a trivial fraction of the sample variance of r_t can be attributed to the measured variation in $E_{t-1}(\tilde{r}_t)$ as a function of $s(\eta_t)$, the results are consistent with the conclusion that almost all the variation in $\tilde{r}_t$ is a reflection of unanticipated changes in purchasing power. Fama (1975, 1976, chap. 6) comes to the same conclusion, but on the basis of autocorrelations of r_t that are close to zero and correlations between r_t and R_t that are likewise indistinguishable from zero

In the same way, $s(\eta_t)$ is an appropriate measure of uncertainty at $t - 1$ about $E_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$, as long as variation through time in $E_{t-1}(\tilde{r}_t)$ is a small fraction of the variation in R_t From (13), we have

$$R_t - R_{t-1} = [E_{t-1}(\tilde{r}_t) - E_{t-2}(\tilde{r}_{t-1})] - [E_{t-1}(\tilde{\Delta}_t) - E_{t-2}(\tilde{\Delta}_{t-1})].$$

Part B of table 4 shows the standard deviations of $R_t - R_{t-1}$ and $\hat{r}_t - \hat{r}_{t-1}$, where $\hat{r}_t$, the estimate of $E_{t-1}(\tilde{r}_t)$, is from the estimated regression function of r_t on $s(\eta_t)$ for the 1953–71 period [5] The estimate of $s(\hat{r}_t - \hat{r}_{t-1})$ is 000033 while the standard deviation of the change in the spot rate is 000325 Since the ratio of variances is 01, the measured changes in $E_{t-1}(\tilde{r}_t)$ are trivial relative to the variation through time of R_t Thus the assumption that the expected real return on a 1-month bill is constant through time is not literally correct, but the interesting implication of the assumption, that variation in R_t mirrors variation in $E_{t-1}(\tilde{\Delta}_t)$, seems to be a good approximation to the world This is also the conclusion drawn by Fama (1975) on the basis of regressions of Δ_t on R_t that yield slope coefficients indistinguishable from -1 0

The Components of the Uncertainty in the Real Returns on Bills

Table 4 summarizes the behavior through time of estimates of the expected premiums in the returns on 2- and 3-month bills obtained from the estimated regression functions in table 2 for the January 1953–July 1971 period The standard deviation of the month to month changes in $\widehat{H_{3t} - R_t}$, 000032, is similar to the standard deviation of the changes in $\hat{r}_t$, while the standard deviation of the changes in $\widehat{H_{2t} - R_t}$, 000018, is smaller Thus, if the measured uncertainty of changes in $E_{t-1}(\tilde{r}_t)$ is small relative to uncertainty about future expected inflation rates, then uncertainty about future expected inflation rates is also large relative to uncertainty about changes in expected premiums

In sum, equation (17) says that each future period in the life of a bill contributes three terms to the uncertainty in the bill's premium (1) the change from $t - 1$ to t in the expected premium on the bill for the future period, (2) the change from $t - 1$ to t in the expected real return on a 1-month bill for the future period, and (3) the change from $t - 1$ to t in the expected rate of change in purchasing power for the future period If one takes the regression function estimates at face value, the evidence in table 4 is that the uncertainty of the premium is largely the uncertainty at $t - 1$ about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$. We earlier argued that, in terms of equation (14), the premium itself is a smaller source of uncertainty in the real return on a bill than the uncertainty of the real return on a 1-month bill Since all the uncertainty of $\tilde{r}_t$ is uncertainty about $\tilde{\Delta}_t$, the evidence points to the conclusion that the uncertainties in the real returns on Treasury bills are almost entirely inflation related

---

[5] Since the autocorrelations of $\hat{r}_t$ and R_t in Part B of table 4 are close to one, it is appropriate to measure the uncertainty about future values of these variables from their changes rather than from their levels This also applies to the variables analyzed in the next section of the text

IX. Conclusions

Summary

The goal of the paper is to identify relationships between expected real returns or expected premiums and the two sources of inflation uncertainty in the real returns and premiums on bills The attempts to identify variation through time in expected real returns as a function of the major source of uncertainty in real returns, the uncertainty of $\tilde{\Delta}_t$, are completely unsuccessful We argue, however, that the uncertainty of $\tilde{\Delta}_t$ doesn't seem to change much during the sampling period, and that this might explain why relationships are not observed

Variation through time in uncertainty about future assessments of expected rates of change in purchasing power seems somewhat more systematic, and, perhaps as a consequence, we are more successful in identifying relationships between this source of inflation uncertainty and expected real returns and premiums The expected real return on a 1-month bill is negatively related to $s(\eta_t)$, the measure of uncertainty at $t - 1$ about $\tilde{E}_t(\tilde{\Delta}_{t+\tau})$, $\tau \geq 1$ Expected premiums are positively related to $s(\eta_t)$, with regression coefficients that are larger the longer the maturity of the bill

Some Perspective from Portfolio Theory

As mentioned earlier, the search for relationships between the expected returns on Treasury bills and the inflation-related uncertainties in the returns is consistent with the early term structure literature, where the risks of bills and bonds are equated with the uncertainties in their returns The approach seems inconsistent with modern portfolio theory, where the risk of a security depends on how it contributes to the uncertainties in the returns on diversified portfolios It seems appropriate at this point to look at what we have done from the viewpoint of portfolio theory[6]

In the models of capital market equilibrium based on portfolio theory, there is a linear relationship between the equilibrium expected return on any security and the risk of the security in the market portfolio. The market portfolio, M, is the portfolio of all available securities, with each security weighted in proportion to the market value of all its outstanding units. The risk of any security j in M is $\text{cov}(\tilde{r}_{jt}, \tilde{r}_{Mt})$, the covariance between the real return on the security and the real return on M To test whether the pricing of Treasury bills conforms to the predictions of these models, one might estimate $E(\tilde{r}_{jt})$ and $\text{cov}(\tilde{r}_{jt}, \tilde{r}_{Mt})$ for bills of different maturities and then test whether the differences in estimates of expected returns correspond appropriately to differences in estimates of risk

---

[6] Detailed discussion of portfolio theory, derivative models of capital market equilibrium, and the empirical tests of these models are in Jensen (1972) or Fama (1976, chaps 7–9)

There are problems in implementing this direct approach [7] For example, since variances of bill returns are trivial relative to variances of returns on other securities like common stocks, and since $\text{cov}(\tilde{r}_{jt}, \tilde{r}_{Mt}) \leq \sigma(\tilde{r}_{jt})\sigma(\tilde{r}_{Mt})$, bills are low-risk instruments, and the differences between the values of $\text{cov}(\tilde{r}_{jt}, \tilde{r}_{Mt})$ for different bills are likely to be small In this situation, estimates of $\text{cov}(\tilde{r}_{jt}, \tilde{r}_{Mt})$ are likely to be sensitive to the choice of M (In empirical work a proxy must inevitably be used) It would be important to have a version of M that covered the full range of investment assets (for example, bonds, stocks, and real estate) with appropriate weights assigned to each asset Such a version of M is not yet available and would be difficult to obtain In contrast, these measurement problems are not so acute for common stocks because the ranges of average return and risk covered by stocks are much broader than for bills

Although the approach taken here is not to test directly the implications of portfolio theory for the pricing of Treasury bills, the results obtained have a bearing on any future tests that might try the portfolio approach There is reliable evidence that expected real returns on longer-term bills are higher than on shorter-term bills Average premiums are positive, large relative to their standard errors, and an increasing function of maturity Thus if in fact the pricing of securities follows portfolio theory, the risks of longer-term bills, as measured by $\text{cov}(\tilde{r}_{jt}, \tilde{r}_{Mt})$, must be larger than the risks of shorter-term bills Since the differences in average returns by maturity are small in absolute terms, however, the differences in risk must also be small

Moreover, the results contain information about the source of any higher risks of longer-term bills In intuitive terms, in the models of market equilibrium based on portfolio theory, a security's risk depends on how the uncertainty in its return is related to uncertainties in the return on the market portfolio The evidence suggests that the uncertainty in the premiums in the returns on longer-term bills is almost entirely uncertainty about changes in assessments of future expected inflation rates Thus if the risks of longer-term bonds, as measured by $\text{cov}(\tilde{r}_{jt}, \tilde{r}_{Mt})$, are larger than the risks of shorter-term bonds, the differences in risk must somehow be associated with uncertainty about future expected inflation rates In this situation, uncertainty about future expected inflation rates must be one of the sources of uncertainty in the return on the market portfolio. Given that bills are low-risk instruments, however, uncertainty about future expected inflation rates is certainly a small part of the uncertainty in the return in *M*

---

[7] See Roll (1971)

References

Fama, E F "Short-Term Interest Rates as Predictors of Inflation." *A.E R.* 65 (June 1975) 269–82

——— *Portfolio Decisions and Security Prices* New York Basic Books, 1976.

Fama, E. F., and MacBeth, J "Long-Term Growth in a Short-Term Market " *J. Finance* 29 (June 1974). 857–85

Hicks, J R *Value and Capital* 2d ed. Oxford. Clarendon Press, 1946

Jensen, M "The Foundations and Current State of Capital Market Theory." *Bell J Econ. and Management Sci* 3 (Autumn 1972) 357–98.

Kessel, R A *The Cyclical Behavior of the Term Structure of Interest Rates* New York Nat Bur. Econ Res, 1965

McCulloch, J. H "An Estimate of the Liquidity Premium " *J P E* 83, no 2. (February 1975) 95–119.

Markowitz, H *Portfolio Selection Efficient Diversification of Investment* New York Wiley, 1959

Roll, R "Investment Diversification and Bond Maturity " *J Finance* 26 (March 1971) 51–66.

Samuelson, P. A "Proof That Properly Anticipated Prices Fluctuate Randomly " *Indus Management Rev* 6 (Spring 1965) 41–49

Human Capital and Labor Supply: A Synthesis

Alan S. Blinder

Princeton University

Yoram Weiss

Tel-Aviv University and National Bureau of Economic Research

The joint determination of work and investment in human capital over the life cycle is analyzed. At low rates of impatience investment is decreasing throughout life, as in simpler models which assume hours of work to be fixed. The demand for leisure over the life cycle is "U shaped." Wages rise to a single peak which occurs after the peak in hours of work. Distinctly different patterns arise when the rate of impatience is high. Such individuals may prefer an increasing hours of work profile, and schooling need not be concentrated at the beginning of life. Conditions are provided to determine a critical level of time preference which is sufficient to induce a "normal" life-cycle pattern for investment and work.

1. Introduction

It is by now widely recognized that investment decisions play a major role in the determination of individual age-earnings profiles. Several studies, most notably Mincer (1974), rely on investment in human capital as a leading single hypothesis from which many of the observed regularities in earnings data are derived.

This is a condensed version of a working paper by the same title which contains many of the proofs and derivations omitted here. The longer version is available on request from the Industrial Relations Section, Princeton University, Princeton, New Jersey. The authors have benefited from seminar presentations at the University of Minnesota, Michigan State University, and the National Bureau of Economic Research, West, and from comments by Mark Killingsworth and John Driffill. Blinder's participation in this project has been supported by the National Science Foundation. Weiss's participation has been supported by the Industrial Relations Section, Princeton University, and by the National Bureau of Economic Research.

[*Journal of Political Economy*, 1976, vol. 84, no. 3]
© 1976 by The University of Chicago. All rights reserved.

The investment hypothesis, in its simplest form, can be summarized by the equation[1]

$$Y = F(K, K), \tag{1}$$

where Y is observed earnings, K is potential earnings ("human capital"), and K is its rate of change Competitive equilibrium in the labor market places restrictions on the signs of the partial derivatives Specifically, as long as human capital contributes to earnings (i e , $\partial F/\partial K > 0$), the individual must give up current earnings in order to enhance future earning opportunities (i e , $\partial F/\partial K < 0$) It is typically assumed that at each point in his life the individual faces a spectrum of earnings-investment combinations ("jobs") and selects the one which is optimal

Equation (1) highlights the formal similarity to the investment problem of the firm in the presence of adjustment costs (see Eisner and Strotz 1963) However, the fact that human capital is embodied in the individual and cannot be sold leads to several important differences First, since the property rights to human capital cannot be transferred, the finiteness of life plays a central role in human investment Second, since either the utilization or formation of human capital requires the sacrifice of leisure (a specific consumption good), it is not possible, in general, to separate the consumption and investment problems as is done in the theory of the firm Third, neither human capital nor the investment therein are normally observable, so a human-capital model is useful only if it generates testable predictions about observable variables like wage rates and earnings Finally, lack of data on human capital makes it hard to distinguish empirically among the effects of investment and a myriad of other influences on actual earnings The most obvious of these, of course, is the choice of work intensity, which affects both current earnings (labor-leisure choices) and the rate of human-capital accumulation (training-leisure choices)

The purpose of this paper is to present a simple life-cycle model of investment in human capital in which leisure choices are explicitly incorporated In so doing, we integrate two previously disparate branches of life-cycle theory models of labor supply with exogenous wages[2] and models of human-capital formation with exogenous leisure [3] Of course, to accomplish this we must posit utility maximization as the individual's goal rather than income maximization

---

[1] This precise representation of the model is due to Rosen (1973), but similar formulations are implicit in Becker (1964) and Ben-Porath (1967)

[2] See, for example, Weiss (1972), Blinder (1974, chap 3), and Heckman (1974) The paper by Weiss allows for endogeneity of wages due to learning by doing but not for human investment

[3] See, for example, Ben-Porath (1967), Weiszacker (1967), and Sheshinski (1968) Formally, our approach also embraces life-cycle consumption theory, but we take pains to separate this from the other two problems and have little to say about consumption

Apart from the direct interest in the interaction of labor supply and human investment over the life cycle, such a model is needed to test the robustness of the widely used wealth-maximization models of human-capital accumulation. For example, a standard implication of these models is that a period of specialization in investment (interpreted as schooling), if it exists, will occur only at the beginning of life[4] However, one can imagine that when schooling involves foregone leisure, the existence of pure time preference might lead a utility-maximizing individual to postpone his education Such possibilities appear explicitly in our model Another important implication of the wealth-maximization model is that the fraction of time spent investing falls throughout the postschooling investment period This turns out to be generally true in our model as well, though some exceptions are noted

To our knowledge, there have been four previous attempts to integrate human capital and labor supply as we do here The treatment by Ghez and Becker (1975) is the most general and explores the widest variety of issues But it is also the least ambitious, in that they generally content themselves with stating and interpreting the first-order conditions Our model can be viewed as a special case of theirs, but a case which is pushed much farther The three other studies[5] at least attempt to analyze the shape of the optimal plan and naturally adopt simplifying assumptions in order to do so Typically, the rate of investment in human capital and the supply of labor are related to some key variable such as the stock of human capital or its shadow price (both unobserved) However, since the models generally do not fully determine the behavior of these endogenous variables, a disturbing and unnecessary amount of ambiguity remains Further, except for Heckman (1976), virtually no attention is paid to periods of specialization such as schooling or retirement This seems a major omission

The model presented here is more general than previous work on the subject (with the exception of Ghez and Becker) and yet generates many more concrete conclusions Among the restrictions which our model places on optimal plans are (*a*) Several distinct patterns in investment, work, and leisure may arise, depending on the subjective rate of impatience (i e , the discount rate for future utilities) (*b*) We consider the case where the rate of impatience is "small" (in a sense to be defined later) to be the leading case In this case, specialization in schooling can occur only at the beginning of life, and retirement can come only at the end (*c*) Schooling is followed by a period of on-the-job training (OJT), during which the fraction of potential earnings and time devoted to human-capital formation declines monotonically. (*d*) Investment reaches zero some finite time before retirement (or death, if there is no retirement). Thus there is a finite interval of "pure work" with no investment late in life. This econometric finding of Mincer (1974) is explicitly ruled out by the Ben-Porath (1967) formulation of the problem (which is followed by Heckman [1976] and by Stafford and Stephan [1973]). (*e*) The demand for leisure over the life cycle is "U-shaped" with a tendency to decline during schooling and the early part of OJT and thereafter to rise. (*f*) Wages rise to a single peak, which occurs after the peak in hours of work.

---

[4] On this, see Weiss (1971), Ishikawa (1973), and also the references cited in the preceding footnote

[5] See Landsberger and Passy (1973), Stafford and Stephan (1973), and Heckman (1976)

Of course, we do not pull these rabbits out of the proverbial hat. Like other investigators, we have to make some simplifying assumptions, of which two seem most crucial. First, (1) is given the special form

$$Y = F(K, \dot{K}) = KG(\dot{K}/K),\ G' < 0, \tag{2}$$

so that $G(\cdot)$ is the fraction of potential earnings which is actually realized. The nature of this important function is described briefly in the next section. Here we only wish to point out that the formulation is quite similar to the early work of Rosen (1973) who dealt with the special case $F(K, \dot{K}) = K - C(\dot{K})$, $C'(\dot{K}) > 0$. In both specifications, K is defined as earnings capacity, that is, the amount the individual would earn in the absence of investment. While Rosen orders jobs by their absolute rate of increase of earning potential ($\dot{K}$) and deducts costs additively, we order jobs by their proportionate rate of growth in potential earnings ($\dot{K}/K$) and deduct costs multiplicatively.[6] In principle, either representation is as good as the other, but the multiplicative version leads more naturally to the logarithmic wage functions encountered so frequently in empirical work (Weiss 1975).

Our second important assumption is that, for a given input of time, the quantity of human capital created is proportional to the stock of human capital. This specification, which is suggested by Mincer's work (1974), simplifies the mathematics considerably. Together, these two assumptions enable us to portray the rather complicated optimal control problem on a two-dimensional phase diagram.

2. The Earnings-Investment Frontier

Basic to any model of human capital is the market constraint delimiting the combinations of current earnings and human-capital formation which are obtainable from a given stock of human capital. This function, which we call the "earnings-investment frontier," is defined implicitly by (1). It is clearly downward sloping. Our special form (2) enables us to normalize

---

[6] In his later work, Rosen (1975) also adopts the multiplicative form.

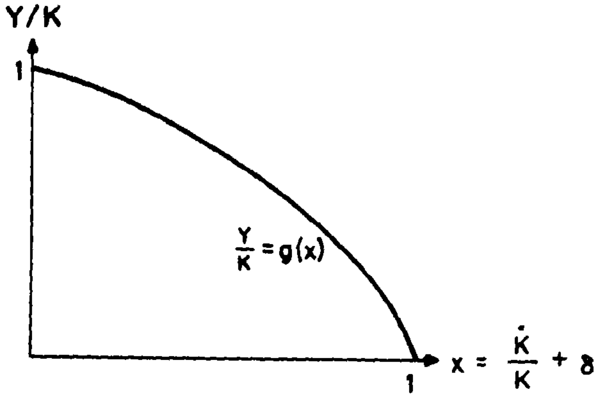


FIG 1—The earnings-investment frontier

this relation so that a single frontier, applicable to any K, can be drawn, thus it plays for us the role that constant returns to scale plays in growth theory The normalized frontier is sketched in figure 1, where the new symbol, x, is defined as $\dot{K}/K + \delta$, where δ is the rate of depreciation of earning capacity As indicated in the introduction, we index jobs by their proportionate rate of growth, x is this index, and it runs from zero to unity Then $x = 1$ indicates the job that yields the maximum feasible rate of growth It makes sense to call this "going to school" and to specify that $g(1) = 0$—that is, that all earnings are sacrificed during schooling Conversely, $x = 0$ is the job where potential earnings are fully realized, so $g(0) = 1$

Throughout the analysis we shall assume that $g(x)$ is a continuous, decreasing, and concave function of x Since $g(x)$ is really a description of an equilibrium wage structure, the assumption of concavity requires some justification

Strict convexity is ruled out by allowing workers to divide their time between jobs or between work and school without incurring significant transactions costs This leaves open the possibility that the frontier might be linear, which would imply that combining training and work on the job is no better than dividing one's time between working and attending school Though possible, we find this notion somewhat unrealistic The essence of learning from experience, as distinct from formal schooling, is that the work environment facilitates the acquisition of productive skills Accordingly, we shall focus on the case in which $g(x)$ is strictly concave, that is, $g''(x) < 0$ [7] However, an almost equivalent model can be developed under the alternative assumption that $g(x)$ is linear

---

[7] For a more detailed derivation of the earnings-investment frontier as a general equilibrium problem, see Rosen (1972)

3. Statement of the Problem

The individual is assumed to derive utility from three sources: the stream of lifetime real consumption, $c(t)$, the fraction of time devoted to leisure, $l(t)$, and the bequest, or terminal value of real (nonhuman) assets, $A(T)$.[8] Here t denotes the individual's age and runs from zero to T, the length of life which is assumed known and exogenous. Specifically, lifetime utility is assumed to be additively separable with a constant rate of time discounting,

$$\int_0^T U(c, l)e^{-\rho t}\,dt + B[A(T)], \tag{3}$$

where $U(c, l)$ and $B[A(T)]$ are assumed to be twice-differentiable, strictly concave functions of their arguments, and ρ is what we call the rate of impatience.[9] We rule out optimal paths with segments of zero consumption or zero leisure by assuming

$$\lim_{c\to 0} U_c(c, l) = \infty \text{ for all } l,$$

$$\lim_{l\to 0} U_l(c, l) = \infty \text{ for all } c.$$

The further stipulation that $U_l(c, 1) > 0$ for any c allows us to consider retirement $[l(t) = 1]$ as an endogenous decision.

What are the constraints on this maximization problem? Letting $h(t)$ denote the fraction of time devoted to market activity (including work and education), the time budget requires

$$h(t) + l(t) = 1 \tag{4}$$

$$h(t) \geq 0 \tag{5}$$

As noted above, the occupational index is bounded between zero and unity

$$0 \leq x(t) \leq 1 \tag{6}$$

Nonhuman capital is generated by the differential equation

$$\dot{A} = rA + g(x)hK - c, \tag{7}$$

where r is the real rate of interest.[10] And human capital is generated by the human-capital production function $\dot{K} = \Phi(x, h, K)$, which can be simplified to

---

[8] It will be assumed that a person's human wealth dies with him, that is, that it cannot be bequeathed.

[9] A variable rate of impatience could be accommodated without too much difficulty. But it would make the notation more cumbersome, and qualitative results would depend on the time profile of $\rho(t)$.

[10] Letting r denote the real rate of interest implicitly incorporates changes in the price of consumer goods.

$$\dot{K} = (axh - \delta)K \tag{8}$$

by one assumption and a convenient definition. (In this expression, a and δ are constants)

The assumption was mentioned earlier that the human-capital production function is homogeneous of degree one in K, namely, $\Phi(x, h, k) = \phi(x, h)K$ The convenient definition is to measure the occupational index, x, in terms of equivalent schooling time, so that working h hours at job x is equivalent (in the production of human capital) to spending xh hours in school[11]

There are two initial conditions corresponding to the individual's endowment of financial and human wealth,[12] $A(0) = A_0 \geq 0$ and $K(0) = K_0 > 0$, and both terminal stocks are to be chosen optimally Under the assumption that

$$\lim_{A_T \to 0} B'(A_T) = \infty,$$

the optimal bequest will be strictly positive

To set up this problem in a form suitable for the application of Pontryagin's maximum principle, substitute (4) into (3), define shadow prices $p(t)e^{-\rho t}$ for human capital and $\mu(t)e^{-\rho t}$ for financial wealth, and write the Hamiltonian function $H(h, x, c, K, A, p, \mu) = e^{-\rho t}\{U(c, 1 - h) + pK(axh - \delta) + \mu[rA + g(x)hK - c]\}$

First-order necessary conditions for a maximum are (i) at each instant, $x(t)$, $h(t)$, and $c(t)$ are chosen to maximize H, given K, A, p, and μ and subject to the constraints (5) and (6), (ii) $\partial H/\partial K = -(d/dt)(e^{-\rho t}p)$ for all t, (iii) $\partial H/\partial A = -(d/dt)(e^{-\rho t}\mu)$ for all t, (iv) $K(T)p(T) = 0$, (v) $\mu(T) = B'(A_T)$

As the reader may verify, conditions i–iii can be written

$$U_c(c, l) = \mu \tag{9}$$

$$\left.\begin{array}{lll} U_l(c, l) = \mu K g(x) + apKx & \text{if} & 0 < l < 1 \\ U_l(c, l) \geq \mu K g(x) + apKx & \text{if} & l = 1 \end{array}\right\} \tag{10}$$

$$\left.\begin{array}{lll} h[apK + \mu K g'(x)] = 0 & \text{if} & 0 < x < 1 \\ h[apK + \mu K g'(0)] \leq 0 & \text{if} & x = 0 \\ h[apK + \mu K g'(1)] \geq 0 & \text{if} & x = 1 \end{array}\right\} \tag{11}$$

$$\dot{\mu}/\mu = \rho - r \tag{12}$$

$$\dot{p}/p = \rho + \delta - axh - g(x)h\mu/p \tag{13}$$

---

[11] This convention means $\phi(x, h) = f(xh)$ The further stipulation that $f(\ )$ is linear is innocuous as long as $g(x)$ is concave An equivalent model arises if $f(\ )$ is concave while $g(\ )$ is linear Equation (8) is suggested by Mincer's work because xh is the change in accumulated years of schooling, $\dot{S}$, so (8) integrates to $\log K = aS - \delta t$

[12] In interpreting these, it should be noted that K_0 includes both the initial endowment of education and "ability," or whatever else it is that determines productivity

4. Interpretation of the Optimality Conditions

The optimality conditions, equations (9)–(13), have straightforward economic interpretations Since $p(t)$ and $\mu(t)$ are, respectively, the shadow prices (in current utils) of human and nonhuman capital, $V(t) \equiv [p(t)/\mu(t)]K(t)$ is the value of the human-capital stock in terms of the numeraire good (consumption) This can be verified by substituting this definition into (13), using (8) and (12), and integrating to obtain

$$V(t) = \int_t^T e^{-r(\tau-t)}h(\tau)g[x(\tau)]K(\tau)\,d\tau, \qquad (14)$$

which states that $V(t)$ is the present value, in time t dollars, of future earnings along the optimal path—a very natural interpretation of the value of the human-capital stock [13]

Using this variable, we can rewrite the optimality condition (11) in a form which is readily interpretable

$$\left.\begin{array}{lll} -g'(x)K = aV & \text{if} & 0 < x < 1 \\ -g'(0)K \geq aV & \text{if} & x = 0 \\ -g'(1)K \leq aV & \text{if} & x = 1 \end{array}\right\} \qquad (15)$$

Here $-g'(x)K$ is the marginal cost in terms of foregone earnings per hour of raising x, and it can be shown that aV measures the marginal benefits in higher future earnings [14]

Condition (15) can also be given the usual rate-of-return interpretation Let π be the rate of interest that equates the discounted benefits to the costs, that is, π is implicitly defined by

$$-g'(x)K = a\int_t^T e^{-\pi(\tau-t)}h(\tau)g(x)K(\tau)\,d\tau$$

as the marginal internal rate of return Then an equivalent statement of (15) is $\pi = r$ if $0 < x < 1$, $\pi \leq r$ if $x = 0$, $\pi \geq r$ if $x = 1$ Thus, during the schooling period ($x = 1$), the internal rate of return on human investment is strictly greater than the interest rate, despite the fact that hours of study are variable [15] During the postschool investment period, the marginal rate of return is always equal to the rate of interest Since these results hold for any arbitrary leisure profile, they are independent of tastes Therefore, one can verify ex post whether individuals have behaved optimally by computing the appropriate rates of return Of course, when labor supply is a choice variable, the rate of return is ill-defined as an ex ante concept

---

[13] The integration that yields (14) requires the terminal condition $V(T) = 0$ This follows from the transversality condition (iv) since human capital cannot be cashed in at the end of life—i e., $K(T) > 0$

[14] The proof can be found in our longer working paper

[15] Equality holds at the instant of leaving school

We now turn to the conditions for optimal work effort (10) When there is no investment ($x = 0$), wages are exogenous to the individual,[16] and, in view of (9), (10) simply states that the marginal rate of substitution (MRS) is equal to the (potential and actual) wage Analysis of the age-hours profile is exactly as in our previous models of dynamic labor supply with exogenous wages Ghez and Becker (1975) have claimed that this marginal condition also holds when wages are endogenous, and that therefore for the analysis of time-versus-goods substitution in consumption we need not worry about why wages change Two remarks can be made about this finding First, when there is investment, the observed wage and the potential wage are very different Nor is their behavior over time identical during OJT the ratio of observed wage/potential wage is monotonically rising Second, the MRS is equated to the potential wage only in the special case of a linear earnings-investment frontier To see this, divide the left-hand side of (10) by $U_c(c, l)$ and the right-hand side by μ (they are equal by [9]) to get $U_l/Uc = a\,(p/\mu)\,Kx + Kg(x) = aVx + Kg(x)$ Then substitute for aV from (15) (assuming an interior solution) to get

$$\mathrm{MRS} \equiv \frac{U_l}{Uc} = K[g(x) - xg'(x)] \tag{16}$$

It is easy to verify that the function in square brackets is equal to one for all x only in the case $g(x) = 1 - x$ Otherwise, it is always strictly greater than one (as long as $0 < x < 1$) We call the right-hand side of (16) the "full wage" Clearly, the first component, $g(x)K$, measures the current benefits per hour of work That the second component, $-xg'(x)K$, measures the future benefits per hour of today's work follows from (15) This second component of the "full wage" prices x at its marginal (not average) value in computing the future benefits The gap between the full wage and the potential wage represents the "pure profits" from OJT

5. The Temporal Succession of Life-Cycle Phases

Due to the possibility of corner solutions, four qualitatively distinct phases might occur in an individual's life cycle Whether each phase occurs, how long each phase lasts, and the temporal order of the phases form the subject of this section We give the four phases descriptive names and numbers which seem to indicate a sensible ordering as follows Phase I, schooling $x = 1$, $h > 0$, Phase II, OJT $0 < x < 1$, $h > 0$, Phase III, Work $x = 0$, $h > 0$, Phase IV, Retirement $h = 0$ Note that when $h = 0$ the value of x is arbitrary, as is clear from (11)

---

[16] Of course, part of the problem is to determine when the individual will opt for leaving his wage exogenous

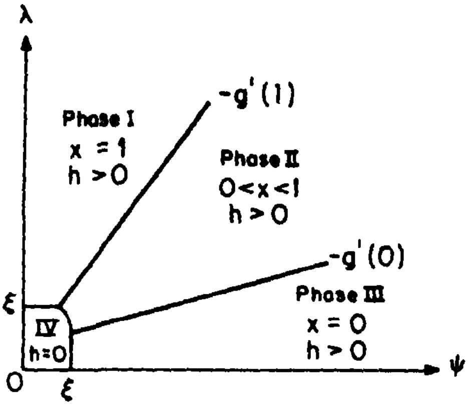


FIG 2—The four phases in (λ, ψ)-space

Economically, this says that a retired person can be considered as holding any job at all (and working zero hours)

To facilitate the exposition, we define two new state variables $\lambda(t) \equiv ap(t)K(t)$ = the value (in utils) of human capital, multiplied by a, and $\psi(t) \equiv \mu(t)K(t)$ = the potential wage converted to utils, and adopt the assumption that the utility function is separable, namely, $U(c, l) = u(c) + v(l)$ Due to our simplifying assumptions, a phase diagram can be drawn in (λ, ψ) space, in which the four phases are defined by (see fig 2)

Phase I (schooling)

$$\frac{\lambda}{\psi} > -g'(1) \tag{17 1}$$

$$v'(1 - h) = \lambda \tag{18 1}$$

Phase II (OJT)

$$\frac{\lambda}{\psi} = -g'(x) \tag{17 2}$$

$$v'(1 - h) = \lambda x + \psi g(x) \tag{18 2}$$

Phase III (work)

$$\frac{\lambda}{\psi} < -g'(0) \tag{17 3}$$

$$v'(1 - h) = \psi \tag{18 3}$$

Phase IV (retirement)

$$v'(1) \geq \lambda x + \psi g(x) \quad \text{for all} \quad x \varepsilon [0, 1] \tag{18 4}$$

It is clear from equations (17) that the precise shape of $g(x)$ strongly colors the likelihood and length of each phase. For example, if $-g'(1) = \infty$, then the individual will never specialize in schooling The reason is that there are jobs (x's very near unity) which are nearly as effective as schooling in imparting skills, but which pay a positive wage Since finiteness of $-g'(1)$ is a minimal requirement for schooling to be possible, we assume it to be true.

As another extreme, if $g'(0) = 0$, then Phase III (work without training) is impossible because there would be a job with a wage arbitrarily close to the no-training job, but which gave a finite amount of training

In this space, the optimal trajectory can, in principle, begin anywhere But, since the transversality condition is $p(T) = 0$, it must terminate on the horizontal axis This trivial observation already establishes that anyone who ever works will indeed have a Phase III—that is, a finite period of work without OJT In other words, as long as we do not rule Phase III out of court by assuming $-g'(0) = 0$, Phase III must be the last phase before retirement (or death) in any optimal plan We believe this makes sense and is an argument against using the Ben-Porath (1967) specification

One other interesting observation can be made already If ψ, the potential wage, is less than $\xi \equiv v'(1)$, the marginal utility of complete leisure, then the work activity is unattractive Similarly, if $\lambda < \xi$, the schooling activity is unattractive However, figure 2 shows that there is a region where both of these inequalities hold, and yet the individual chooses not to retire This means that, for some values of ψ and ξ, the option of combining work and schooling by OJT keeps the individual on the job when he would otherwise retire It can be shown that this is a direct implication of strict concavity of the earning-investment frontier [17]

To determine the nature of optimal trajectories in figure 2, it is necessary to know how λ and ψ change over time Using the definitions of λ and ψ and equations (8), (12), and (13), it follows that

$$\dot{\lambda} = \rho\lambda - a\psi g(x)h \tag{19}$$

$$\dot{\psi} = (ahx - \gamma)\psi, \tag{20}$$

where $\gamma \equiv r - \rho + \delta$

The sign of γ is a major determinant of the qualitative behavior of the model To save space, we present a detailed analysis only of the case $\gamma > 0$, which is probably more realistic (being consistent with an

---

[17] The northeast boundary of Phase IV is concave because it is defined by

$$Q(\lambda, \psi) \equiv \max_{0 \leq x \leq 1} [\lambda x + \psi g(x)] = \xi$$

In the case $g(x) = 1 - x$, the retirement region expands to fill the entire square increasing consumption profile)[18] and also more interesting analytically As we shall see, a complication arises from a phenomenon which we call "cycling," that is, the recurrence of a specific phase more than once in the optimal life cycle It turns out that there is little room for such cycling in the case $\gamma \leq 0$.

Let us, therefore, first deal briefly with people with "high impatience," that is, with $\rho > r + \delta$[19] The following propositions, which we state without proof, rather sharply delimit the kinds of life cycles that could ever be optimal for such a person. (*a*) The only possibility of cycling is that an optimal path might include two disjoint periods of work separated by a period of OJT (*b*) If there is a period of schooling, it comes either at the beginning of life or immediately after retirement. A period of OJT follows schooling (*c*) If there is a period of retirement, it comes at the beginning of life (*d*) If there is a period of OJT (and there will be one unless the optimal path includes no training whatever), it is followed by a period of work (*e*) The last years of life are a period of work.

Thus the typical life cycle for persons with high impatience, assuming most people find it optimal to take some schooling and some retirement, is roughly as follows an initial period of retirement is followed by schooling, then by OJT, and then by pure work until death.

Of course, life cycles with early retirement are rarely observed in practice. But this is not because such programs are irrational Individuals with very high impatience (or very high positive exogenous wage growth) will want to bunch their leisure early in life To do so they will have to work very hard when they are old, since consumption depends on lifetime discounted earnings We may surmise that it is the absence of perfect capital markets that precludes all but inheritors of large fortunes from pursuing such a program

Hereafter we focus on what we take to be the leading case $\gamma > 0$ In order to analyze the behavior of λ and ψ over time, it is necessary to locate the stationary loci, $\dot{\lambda} = 0$ and $\dot{\psi} = 0$, in the four regions

Because $\rho > 0$, equation (19) implies that the $\dot{\lambda} = 0$ locus can never enter Phase I ($x = 1$) or Phase IV ($h = 0$), for then $\dot{\lambda}$ would be positive In Phase III ($x = 0$), the $\dot{\lambda} = 0$ locus is $\lambda = (a/\rho)\psi h$, where h is implicitly defined as a function of ψ by (18.3) It can be shown (proof omitted) that the $\dot{\lambda} = 0$ line begins at the point $\lambda = 0$, $\psi = \zeta$ and slopes upward until it intersects the $-g'(0)$ ray at point A in figure 3

Analogously, when $\gamma > 0$ the $\dot{\psi} = 0$ locus cannot enter Phase III ($x = 0$) or Phase IV ($h = 0$), for then $\dot{\psi}$ would have to be negative In

---

[18] This simple identification of rising or falling consumption paths with $r - \rho$ is true only under separable utility For the more general case, see Weiss (1972) Our point is that this specific model behaves most like the real world when $r > \rho$

[19] Note that if the rate of growth of wages in the absence of investment, $-\delta$, is positive (due to economy-wide productivity changes) and exceeds the real rate of interest, then all persons are in this class.

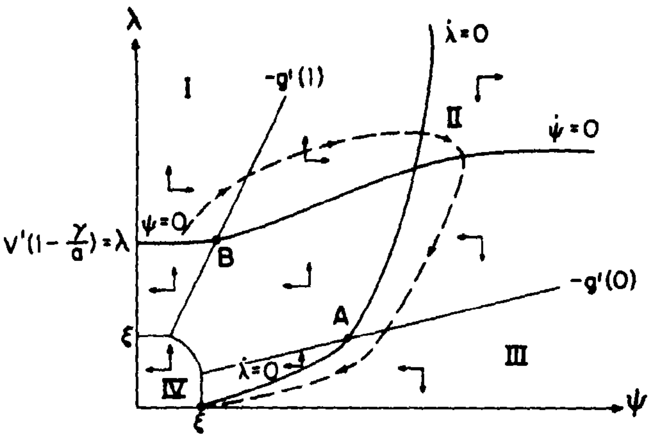


FIG 3—Phase diagram when $\gamma > 0$

Phase I, $\dot{\psi}$ is stationary only when $h = \gamma/a$ Since h is implicitly defined as a function of λ by (18 1), this means that only one specific value of λ makes $\dot{\psi} = 0$ in Phase I (see fig 3)

The positions of the two stationary loci in Phase II can be derived by simultaneously manipulating equations (17 2) and (18 2) As this task is rather arduous and uninteresting, it will not be described here Instead we confine ourselves to stating without proof the following three basic properties (*a*) The $\dot{\lambda} = 0$ locus in Phase II begins at point A in figure 3 and slopes upward from there (*b*) The $\dot{\psi} = 0$ locus in Phase II begins at point B in figure 3 and slopes upward from there (*c*) While both loci must begin with positive slopes, either or both slopes might become negative somewhere in the region in contrast to the way they are depicted in figure 3 This is not important What is important is that the two loci must have a unique intersection in Phase II with the $\dot{\lambda} = 0$ line cutting the $\dot{\psi} = 0$ line from below [20]

6. Labor Supply and Human Investment in a Normal Life Cycle

It can be seen from figure 3 that if the life plan includes schooling and has no cycles, the only possibility is that schooling comes first, followed by OJT, work, and then retirement [21] We therefore call this the "normal" life cycle and proceed to examine its properties

---

[20] The rather lengthy proof of this assertion appears in the appendix to our longer working paper

[21] This is actually not quite so obvious, since the diagram makes it look as though retirement might come first However, we show in Section 7 that the same conditions that preclude cycling also preclude early retirement

Phase I. Schooling

Since λ is rising through time during Phase I, it is clear from (18 1) that the amount of time devoted to schooling rises steadily as education progresses—a prediction to which all former graduate students will doubtless attest The intuition behind this is simply that the value (in utils) of the human-capital stock $\lambda(t)$ rises, making leisure more expensive [22]

We may also make some rough judgments on the concavity of the hours profile, $h(t)$ Differentiating (18 1) logarithmically gives

$$\left[\frac{-v''(l)}{v'(l)}\right] h = \frac{\lambda}{\lambda} = \rho$$

Denote the expression in brackets by $R(h) > 0$ and take the time derivative again to get

$$R'(h)(h)^2 + R(h)h = 0 \qquad (21)$$

It is clear from (21) that h and $R'(h)$ have opposite signs What is the likely sign of $R'(h)$? Noting that the time budget makes $dh/dl = -1$ everywhere, we can write $R'(h)$ as

$$R'(h) = \frac{d}{dl}\frac{v''(l)}{v'(l)}$$

In the case of utility functions for choices involving risk, the ratio $v''(l)/v'(l)$ is called the degree of "absolute risk aversion" (Pratt 1964) and is generally thought to be an increasing function of l (i e , to fall in absolute value as l rises) While the present model does not discuss risk, if our $v(l)$ function also has this property, then $h(t) < 0$—that is, h is a concave function of time during schooling [23] We take this to be the leading case And since $K/K = ah - \delta$ in Phase I, the rate of growth of potential wages would also be an increasing and concave function of time

Phase III Work

Since there are some formal similarities between Phase I and Phase III, it is convenient to take up next the case of work with no training. Here labor supply is governed by the usual condition equation (18 3) states that the marginal rate of substitution between leisure and consumption goods is equated to the real wage Since, by (20), $\dot{\psi}/\psi = -\gamma$ in Phase III, logarithmic differentiation of (18.3) yields

---

[22] Marginal calculations during this phase involve only leisure and schooling The work activity is dominated

[23] If, instead, $(d/dl)[v''(l)/v'(l)] < 0$, h will be a convex function of time

$$\frac{-v''(l)}{v'(l)} h = -\gamma = \rho - r - \delta \tag{22}$$

So people with "normal" time discount rates have diminishing labor supply in Phase III By contrast, people with high impatience offer increasing amounts of labor to the market despite wages which are falling if $\delta > 0$. It cannot be stressed too much that these contrasting behavior patterns have absolutely nothing to do with competing income and substitution effects, although cross-sectional studies of labor supply might possibly confound the two phenomena

From (22) it is clear that the concavity issue is precisely as it was in Phase I The sign of h depends only on the behavior of "absolute risk aversion" as l rises, and the more attractive utility functions imply $h < 0$

The behavior of wage rates (actual and potential coincide) and earnings in Phase III is also of interest Letting W be the observed wage, we have $\dot{W}/W = K/K = -\delta$, a constant So logarithmic age-wage profiles should be straight lines in Phase III, which are falling, flat, or rising according as δ is greater than, equal to, or less than zero

Finally, consider earnings, $Y(t) = h(t)W(t)$ Since $(Y/Y) = (h/h) + (\dot{W}/W) = (h/h) - \delta$, if $\delta > 0$ (i e , if depreciation outweighs economy-wide wage growth), earnings will surely decline However, if δ is sufficiently negative, they may not But regardless of the slope of the logarithmic earnings profile, it will certainly be concave if $h(t)$ is for

$$\frac{d}{dt}\left(\frac{\dot{Y}}{Y}\right) = \frac{d}{dt}\left(\frac{h}{h}\right) = \frac{hh - (h)^2}{h^2} < 0 \qquad \text{if} \qquad h < 0$$

Phase II On-the-Job-Training

To analyze the behavior of hours of work and investment in Phase II, it is convenient to transform the differential equations so that x and h rather than λ and ψ are the variables In (x, h) space, Phase II is the open unit square, Phase I is the vertical line $x = 1$, Phase III is the vertical line $x = 0$, and Phase IV is the horizontal axis (see fig 4) In a life cycle without cycling, we enter from Phase I and exit to Phase III

The first-order conditions for an interior maximum are (17 2) and (18 2) which, since the Hamiltonian is strictly concave in (x, h) for given (λ, ψ), define x and h as unique functions of λ and ψ [24] Given these, the

---

[24] Uniqueness of λ and ψ for given x and h is obvious since (17 2) and (18 2) are linear in λ and ψ

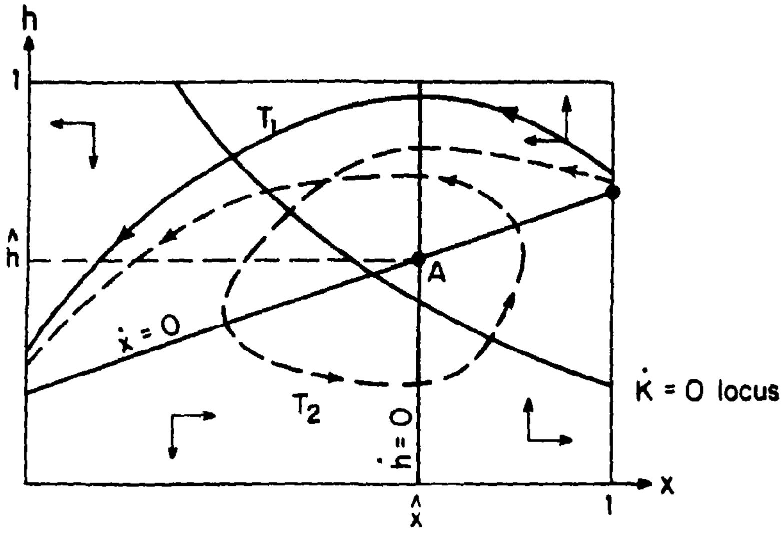


FIG 4—Alternative phase diagram in (x, h)-space

two differential equations in $\dot{\lambda}$ and $\dot{\psi}$ can be transformed into a pair of differential equations in h and x. The results are

$$\frac{-v''(l)}{v'(l)}h = \rho - \frac{r+\delta}{1+\eta(x)} \tag{23}$$

$$\frac{g''(x)}{g'(x)}x = r + \delta - ahx\frac{1+\eta(x)}{\eta(x)}, \tag{24}$$

where $\eta(x) \equiv -xg'(x)/g(x)$ is the (sign-corrected) elasticity of the $g(x)$ function.

Using (23) and (24), it is easy to partition figure 4 into four regions by the $x = 0$ and $h = 0$ loci. The $h = 0$ locus has the simple form

$$\eta(x) = \gamma/\rho, \tag{25}$$

which defines a unique value of x.[25] Call this $\hat{x}$. Using (24), the end points of the $x = 0$ locus, which is defined by

$$r + \delta = ahx\left(\frac{1+\eta(x)}{\eta(x)}\right), \tag{26}$$

---

[25] This is because $\eta(x)$ rises monotonically from $\eta(0) = 0$ to $\eta(1) = +\infty$.

can easily be determined. When $x = 1$, $h = (r + \delta)/a$, and when $x = 0$, $h = -g'(0)(r + \delta)/a < (r + \delta)/a$. It can also be shown that the locus is upward sloping.[26]

It is clear from figure 4, then, that h can be no less than $(r + \delta)/a$ at the start of Phase II[27] And the trajectory must proceed smoothly from $x = 1$ to $x = 0$ with $x < 0$ everywhere, like path T_1 in figure 4, rather than form "mini cycles" like path T_2 in figure 4 This is because (23) and (24) define x and h as functions of x and h only—that is, because x and h do not depend directly on either the state variables or the costate variables It therefore follows that (x, h) must be the same whenever (x, h) are But a path like T_2 would have to cross itself—include two distinct points in time with the same (x, h) and different (x, h)—which is impossible This is an important result, since it shows that within Phase II, $x(t)$ is monotonically declining—a property which holds in the simple Ben-Porath (1967) model and which is vital if human-capital theory is to account for the gross facts It also implies that within Phase II labor supply rises to a single peak and then declines

Nothing can be said in general about the concavity of the $x(t)$ profile However, we can use equation (23) to determine the concavity of $h(t)$ Letting $R(h) \equiv -v''(l)/v'(l)$ as before, its time derivative is

$$R(h)h = \frac{(r + \delta)\eta'(x)}{[1 + \eta(x)]^2}\, x - (h)^2 R'(h) \qquad (27)$$

Since the first term on the right-hand side is negative, $h < 0$ for utility functions with $R'(h) \geq 0$[28]

The behavior of potential wages, $K(t)$, can be displayed conveniently on the (x, h) diagram Since $K/K = ahx - \delta$, the $K = 0$ locus (which exists only if $\delta > 0$) is the rectangular hyberbola, $hx = \delta/a$, shown in figure 4 By showing that the intersection of the $x = 0$ and $h = 0$ loci (labeled point A in fig 4) lies above this hyperbola, we will establish that the peak in labor supply precedes the peak in human capital Point A is defined by $\hat{x}$ and the $\hat{h}$, which satisfies

$$r + \delta = a\hat{h}\hat{x}\left(\frac{1 + \eta(\hat{x})}{\eta(\hat{x})}\right)$$

---

[26] Proof Using the definition of $\eta(x)$ and differentiating (26) yields

$$0 = ax\,\frac{1 + \eta(x)}{\eta(x)}\,\frac{dh}{dx} + ah\,\frac{g(x)g''(x)}{g(x)^2}$$

Since $g''(x) < 0$, we have $dh/dx\,|_{x=0} > 0$

[27] We must assume $a > r + \delta$ if there is to be any training at all

[28] When $R'(h) < 0$, (27) is ambiguous a priori However, we already know that there is a unique maximum (where $\dot{h} = 0$), and it is clear that $\ddot{h} < 0$ in the neighborhood of this maximum

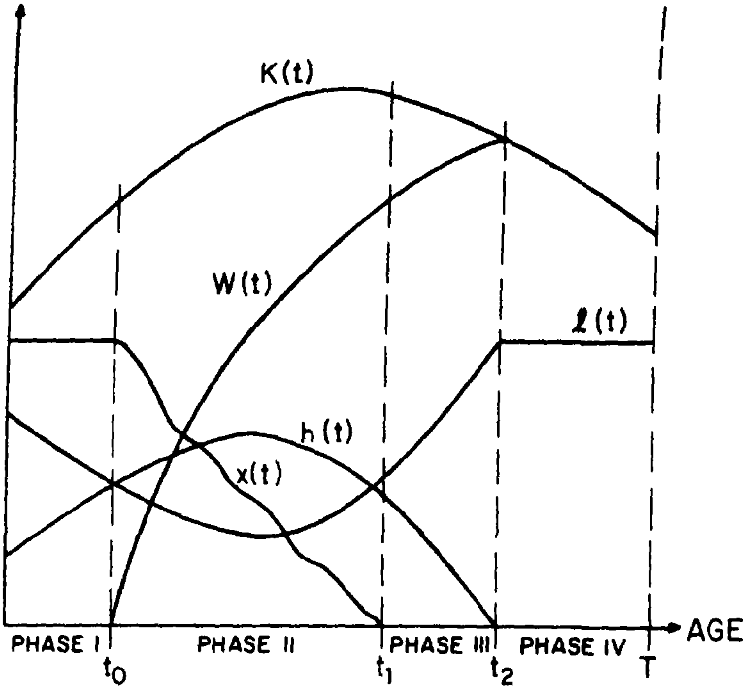


FIG 5—The age profiles of human capital, wages, leisure, work, and investment in a "normal" life cycle

But $\eta(\hat{x}) = \gamma/\rho$, so $\hat{h}\hat{x} = \gamma/a = (r - \rho)/a + \delta/a > \delta/a$ when $r > \rho$ So $r \geq \rho$ is sufficient though certainly not necessary for the peak in labor supply to precede the peak in human capital

What about the peak in observed wages, $W(t) = K(t)g[x(t)]$? Since $\dot{W}(t) = Kg'(x)\dot{x} + g(x)\dot{K}$, and since $g'(x)$ and $\dot{x}$ are negative while $g(x)$ is positive, the peak in observed wages (if any) must follow the peak in potential wages (if any) Of course, if $\delta = 0$, both actual and potential wages reach their peak at the Phase II–Phase III switch point and are level thereafter

Figure 5 summarizes our results It depicts the time profiles of labor supply, human capital, and observed wage rates in the four phases for the case $\gamma > 0$, $\delta > 0$ The peak in observed earnings, $h(t)W(t)$, must come between the peak in h and the peak in W, though its relation to the peak in K is unclear If the $\delta = 0$ case is a good benchmark, earnings peak earlier than human capital

Figure 5 shows wages and hours of work rising together at first (when wages are low), then moving in opposite directions (when wages are higher), and finally falling together (when wages are again low). Once again, we stress that this has nothing whatsoever to do with income and substitution effects—although cross-sectional studies over diverse age cohorts might mistakenly identify this phenomenon as a backward-bending labor-supply function

7. The Question of Cycling

We now address ourselves to two important questions which have thus far been avoided (*a*) When cycling arises, what broad characterizations of the optimal path can be made? (*b*) What meaningful conditions can be derived which exclude the possibility of cycling?

In Section 6, we proved that the optimal trajectory can never cross itself[29] This result enables us to answer the first question, since it implies that cycles must be "expanding" as in figure 6 rather than "contracting " Thus a cycling path can be broken down into several "quasi life cycles," each beginning with a period of schooling (with the possible exception of the first) Since the contours of constant work effort look like blowups of the border of the retirement region (which is the special case $h = 0$), with higher contours connoting higher h, we see that the individual works harder during his second "quasi cycle" than in his first Also, he spends more time in Phase II

We can develop strong conditions which rule out cycles by considering the minimum length of time required to complete a quasi cycle If this time exceeds the available life, T, then cycles are impossible In particular, we focus on the last quasi cycle and let t_0 = age of starting last schooling period, t_1 = age of starting last OJT period, t_2 = age of starting last work period, and t_3 = age of ending last work period (possibly $t_3 = T$) These four points on the optimal trajectory are indicated in figure 6

Consider first the length of the schooling period, $t_1 - t_0$ Since we are considering a cycling path, we know that[30]

$$h(t_0) \leq \gamma/a = \frac{r + \delta - \rho}{a} \tag{28}$$

(see fig 6) Furthermore, we know from figure 4 that

$$h(t_1) \geq (r + \delta)/a \tag{29}$$

Finally, we know that optimal behavior implies that $v'(1 - h)$ grows at the exponential rate ρ in Phase I (see eqq [18 1] and [19]) Thus we can derive a minimal time required to pass through Phase I which will depend on the rate of impatience, the elasticity of the marginal

---

[29] Actually, we only proved that it could never cross itself in Phase II It is trivial to prove this for the other phases as well

[30] Note that no such statement can be made about the schooling period in a trajectory without cycling Thus the bound to be derived will not apply to "normal" life cycles

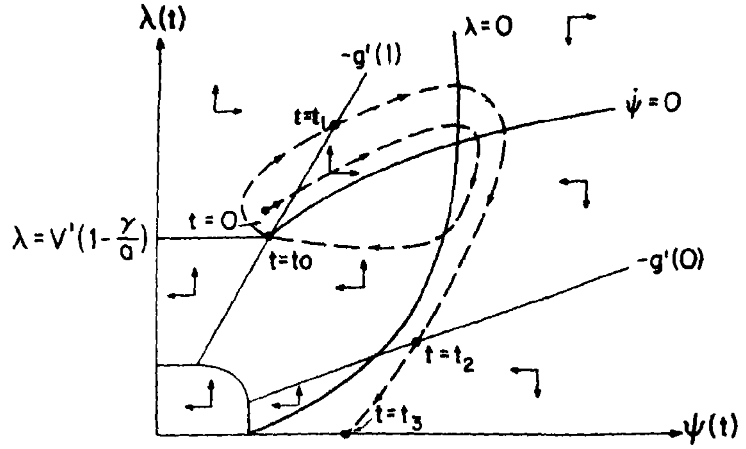


FIG 6—A possible cycle when $\gamma > 0$

utility function, and the parameters a, r, and δ Specifically, since $\lambda(t_0) \leq v'(1 - [\gamma/a])$ by (28), and $\lambda(t_1) \geq v'[1 - (r + \delta)/a]$ by (29), we have

$$e^{\rho(t_1 - t_0)} \geq \frac{v'\left(\frac{a - r - \delta}{a}\right)}{v'\left(\frac{a - r - \delta + \rho}{a}\right)} \qquad (30)$$

or

$$t_1 - t_0 \geq \frac{1}{\rho} \log \frac{v'\left(\frac{a - r - \delta}{a}\right)}{v'\left(\frac{a - r - \delta + \rho}{a}\right)} > -\frac{1}{a} \frac{v''\left(1 - \frac{\gamma}{a}\right)}{v'\left(1 - \frac{\gamma}{a}\right)} \equiv S^* \qquad (31)$$

The inequality in (31) follows if we assume, as in Section 6, that $d/dl\,[v''(l)/v'(l)] > 0$ and provides us with a lower bound, S^*, for the interval $t_1 - t_0$ This lower bound increases as ρ falls

It is also intuitively clear that there must be some minimal period of work during which the benefits from schooling are realized This is the idea of our lower bounds on the lengths of Phases II and III, which depend basically on the trade-offs embodied in the earnings-investment frontier To derive these bounds, we examine the behavior of the ratio $\theta(t) \equiv p(t)/\mu(t)$, which is the price of human capital in terms of consumption goods Using this definition, t_1, t_2, and t_3 are defined by (see fig 6)

$$a\theta(t_1) = -g'(1) \qquad (32)$$

$$a\theta(t_2) = -g'(0) \tag{33}$$

$$\theta(t_3) = 0 \tag{34}$$

It is easy to show that

$$\dot{\theta} = (r + \delta)\theta - h[g(x) - xg'(x)] \text{ in Phase II} \tag{35}$$

$$\dot{\theta} = (r + \delta)\theta - h \text{ in Phase III} \tag{36}$$

First consider Phase II, where the solution to (35) is

$$\theta(t)e^{-(r+\delta)(t-t_1)} = \theta(t_1) - \int_{t_1}^{t} e^{-(r+\delta)(\tau-t_1)}h[g(x) - xg'(x)]\, d\tau$$

Setting $t = t_2$ and using (32) and (33), we have

$$-g'(0)e^{-(r+\delta)(t_2-t_1)} = -g'(1) - a\int_{t_1}^{t_2} e^{-(r+\delta)(\tau-t_1)}h[g(x) - xg'(x)]\, d\tau$$

Now, replacing $h(t)$ in the integral by 1.0 and replacing $g(x) - xg'(x)$ by $-g'(1)$ makes the integral strictly larger, so we have the inequality

$$-g'(1)a\int_{t_1}^{t_2} e^{-(r+\delta)(\tau-t_1)}\, d\tau > -g'(1) + g'(0)e^{-(r+\delta)(t_2-t_1)}$$

or

$$a\,\frac{1 - e^{-(r+\delta)(t_2-t_1)}}{r + \delta} > 1 - \frac{g'(0)}{g'(1)}\,e^{-(r+\delta)(t_2-t_1)}$$

Our lower bound on the length of Phase II, call it J^*, therefore satisfies[31]

$$J^* = \frac{1}{r + \delta}\log\frac{a - \dfrac{g'(0)}{g'(1)}(r + \delta)}{a - r - \delta} \tag{37}$$

The bound depends on the parameters a, r, and δ and also on the degree of concavity of the earnings-investment frontier.[32] Tastes are not involved since we obtained the bound by assuming the individual works as hard as he can.

Almost the same reasoning can be used to place a lower bound, W^*, on the length of the last Phase III. Solving (36) explicitly gives

$$\theta(t)e^{-(r+\delta)(t-t_2)} = \theta(t_2) - \int_{t_2}^{t} e^{-(r+\delta)(\tau-t_2)}h(\tau)\, d\tau$$

---

[31] It should be clear that the bound applies equally well to the OJT period of a normal life cycle, since the only things assumed are that Phase II is preceded by Phase I and precedes Phase III.

[32] That is, on how much smaller $-g'(0)$ is than $-g'(1)$.

Letting $t = t_3$ and using (33) and (34), we obtain

$$-g'(0) = a \int_{t_2}^{t_3} e^{-(r+\delta)(\tau - t_2)} h(\tau)\, d\tau < a \int_{t_2}^{t_3} e^{-(r+\delta)(\tau - t_2)}\, d\tau,$$

since $h(\tau) < 1$. Thus $-g'(0) < [a/(r + \delta)]\,[1 - e^{-(r+\delta)(t_3 - t_2)}]$, or

$$W^* = \frac{1}{r + \delta} \log \frac{a}{a + g'(0)(r + \delta)} \tag{38}$$

Again, the bound depends on a, r, δ, and the $g(x)$ function.

Putting these results together, we find that there are bounds on the minimum lengths of Phases II and III which are independent of ρ, and there is a bound for the minimal schooling period which depends on ρ. Let ρ^* be the value of ρ, satisfying $S^*(\rho^*) = T - J^* - W^*$. Then cycling is certainly impossible for persons with $\rho \leq \rho^*$. Since we know that cycling is only a problem when $\rho < r + \delta$, a sufficient (but far from necessary) condition to rule out cycling is $r + \delta \leq \rho^*$. The reader should note that the condition $\rho \leq \rho^*$ rules out early retirement in the $\gamma > 0$ case as well, since the bounds $S^*(\rho^*)$, J^*, and W^* apply here as well. In a word, we will never get cycles if either ρ is "large" (i.e., $\rho \geq r + \delta$) or ρ is "small" (i.e., $\rho \leq \rho^*$). If $\rho^* < r + \delta$, there will be an intermediate range in which cycling is possible.

It may be of interest to work out a numerical example of these bounds. Suppose a, which would be the rate of return on human capital in the case of an infinite lifetime, is 8 percent, while $r + \delta$ is 6 percent. A simple form for the earnings-investment frontier is the quadratic $g(x) = 1 - \frac{1}{2}x - \frac{1}{2}x^2$. Given these choices, (37) and (38) can be used to compute $J^* = 18.3$, $W^* = 7.8$. From (31) it is clear that a facile choice of utility function is $v'(l) = e^{-l}$. For a rate of impatience of $\rho = .02$, (31) gives $S^* = 12.5$ as the minimal schooling period. Adding these, we find that a new cycle cannot start unless there are more than 38.6 years remaining.[33]

8. In Conclusion

We have presented a life-cycle model of the behavior of a utility-maximizing individual free to allocate his daily time budget among leisure, work, and education. The main substantive conclusions are listed in Section 1 and depicted in figure 5. They need not be rehashed here.

---

[33] The reader is reminded that these are all weak lower bounds. The actual amount of time taken by the last quasi cycle is thus strictly greater (and probably considerably greater) than 38.6 years.

Several generalizations of the analysis immediately come to mind. More general forms of the human-capital production function could be tried, as could a nonseparable utility function. Indeed, the list of arguments of the utility function could be expanded to include human capital or, indeed, nonhuman capital. More important than all these, we imagine, would be allowance for the capital-market imperfections that severely constrain the choices of those poorly endowed with financial wealth.

Still, as long as we interpret the results as a benchmark around which there will surely be deviations (some systematic, some random), there are a number of interesting uses for the model

First, it may be possible to do the usual kinds of comparative-dynamic exercises How would an increase in initial financial wealth alter the optimal plan? What about an increase in initial human wealth? Is the optimal plan continuous with respect to these endowments? How sensitive is the optimal policy to the rate of impatience—a taste parameter that presumably differs across people? The reader can no doubt think of many similar questions While we suspect that most of these can only be answered under specific functional forms for $v(l)$ and $g(x)$, some weak results may hold in greater generality

Second, the model can provide the micro-foundation for a simulation model of the income distribution along the lines of Blinder (1974) The chief conclusion of that work is that the distribution of wage rates, taken there to be exogenous, is the principal contributor to income inequality The present model can generate that distribution endogenously, given assumed distributions of tastes and endowments, and thus can fill what is perhaps the major gap in the positive theory of income distributions.

Third, the long-run incidence of various taxes in a world with human-capital accumulation is virtually unexplored territory [34] It seems feasible to incorporate some simple taxes—such as a linear income or wage tax—into the model and examine, either analytically or through simulation, the effects of these programs on the acquisition of human capital It could be that human investment responds to taxation more substantially than do hours of work

In a word, there are a host of interesting and important questions which simply cannot be addressed by a life-cycle model which considers either labor-leisure choices or labor-education choices but not both [35] By demonstrating the feasibility of handling both decisions together, we hope to have hastened the day when the powerful tools of life-cycle economic theory will be brought to bear on these issues

---

[34] A preliminary discussion is offered by Boskin (1976)

[35] For example, if income maximization is the posited goal, a proportional tax (subsidv) on wages cannot possibly alter behavior

References

Becker, G *Human Capital* New York Nat Bur. Econ. Res , 1964

Ben-Porath, Y "The Production of Human Capital and the Life Cycle of Earnings." *J P E* 75, no 4 (August 1967) 352–65.

Blinder, A S *Toward an Economic Theory of Income Distribution.* Cambridge· M I.T. Press, 1974

Boskin, M "Notes on the Tax Treatment of Human Capital " *Proceedings of the Treasury Conference on Tax Research* Washington, Department of the Treasury, 1976

Eisner, R , and Strotz, R "The Determinants of Business Investment " In *Impacts of Monetary Policy,* edited by the Commission on Money and Credit Englewood Cliffs, N J Prentice-Hall, 1963

Ghez, G , and Becker, G. *The Allocation of Time and Goods over the Life Cycle* New York Nat Bur. Econ Res , 1975

Heckman, J "Estimates of Human Capital Production Functions Embedded in a Life Cycle Model of Labor Supply " In *Household Production and Consumption,* edited by N Terleckyj Studies in Income and Wealth, vol 40 New York Columbia University Press, forthcoming 1976

——— "Life Cycle Consumption and Labor Supply An Explanation of the Relationship between Income and Consumption over the Life Cycle " *A E R* 64 (March 1974) 188–94

Ishikawa, T "A Simple Jevonian Model of Educational Investment Revisited " Discussion Paper no 289, Harvard University, 1973

Landsberger, M , and Passy, U "Human Capital, Its Shadow Price and Labor Supply " Technion Mimeograph Series no 138, Haifa, Israel, 1973

Mincer, J *Schooling, Experience and Earnings* New York Nat Bur Econ Res., 1974

Pratt, J W "Risk Aversion in the Small and in the Large " *Econometrica* 32 (1964) 122–36

Rosen, S "Learning and Experience in the Labor Market." *J Human Resources* 7 (Summer 1972) 326–42

——— "Income Generating Functions and Capital Accumulation " Discussion Paper no 306, Harvard University, 1973

——— "Towards a Theory of Lifetime Earnings " Unpublished paper, University of Rochester, 1975

Sheshinski, E "On the Individual's Lifetime Allocation between Education and Work " *Metroeconomica* 20 (April–June 1968) 42–49

Stafford, F., and Stephan, P "Labor, Leisure and Training over the Life Cycle " Working Paper no 7374-14, University of Michigan, 1973

Weiss, Y "Learning by Doing and Occupational Specialization " *J Econ Theory* 3 (June 1971) 189–98

——— "On the Optimal Pattern of Labor Supply " *Econ J* 82 (December 1972). 1293–1315

——— "The Earnings of Scientists, 1960–1970 Experience, Age and Vintage Effects." Working Paper no 99, Nat Bur Econ Res , 1975

Weiszacker, C C von "Training Policies under Conditions of Technical Progress A Theoretical Treatment " In *Mathematical Models in Educational Planning* Paris Org Econ Cooperation and Development, 1967

An Economic Analysis of Eminent Domain

Patricia Munch

RAND Corporation

A theoretical analysis of land assembly with and without eminent domain concludes that, contrary to traditional assumptions, eminent domain is not necessarily a more efficient institution than the free market for consolidating many contiguous but separately owned parcels into a single ownership unit In practice, prices paid under eminent domain may differ systematically from the "fair market value" standard, depending on court costs of buyer and seller Evidence from urban renewal supports the hypothesis that, due to the structure of court costs, high-valued properties receive more than market value and low-valued properties receive less than market value

Introduction

Eminent domain (ED) is the legal right to acquire property by forced rather than by voluntary exchange When a buyer seeking to acquire a property has the power of ED, he must attempt to negotiate a voluntary sale But if his highest offer is rejected, he may condemn the property, that is, obtain a forced sale at a price determined in a court of law

In the United States, the use of ED is constrained by constitutional provisions at the federal and state level which typically require that private property only be taken for "public use" and only after payment of "just compensation "[1] Enforcement of these constitutional provisions is divided between the legislatures and the judiciary Legislatures—federal and state—have the right to grant ED power They typically

---

I would like to thank the members of my thesis committee, Professors George J Stigler, Gary S Becker, and Ronald H Coase, for their guidance and valuable insights I also benefited from the comments of an anonymous referee

[1] Generalizations on the legal aspects of ED are based on Lewis (1909) and Nichols (1970)

[*Journal of Political Economy*, 1976, vol 84 no 3]
© 1976 by The University of Chicago All rights reserved

grant it to a particular body or for a specific project, leaving to the discretion of the grantee the decision as to which parcels of land to acquire and whether or not to exercise its latent power of condemnation The main role of the judiciary is to determine "just compensation," not to review the legitimacy of the taking of a particular parcel [2] Thus, ED is effectively a reassignment of property rights· the seller is deprived of his right to refuse to sell and constrained in his right to bargain over price However, because of the just-compensation provision, the curtailment of private property rights implied by ED is less than that inherent in the taxing and regulatory powers of government In fact, ED may be interpreted as a limitation on the police power of the state against private property in real estate (Scheiber 1971)

The growth in the use of ED over time suggests, if the survivorship principle may be applied to institutions, that ED reduces some component of costs to at least one party There are two, not mutually exclusive, possibilities (1) ED reduces total costs, permitting a net gain in efficiency, and (2) ED redistributes costs and therefore wealth

If ED were a simple transfer of well-defined property rights, and if transactions costs under both property-right assignments were zero and wealth effects symmetrical, no change in resource allocation would be implied, according to the Coase theorem (Coase 1960) [3] But in the sparse references to ED in the economic literature,[4] the converse is assumed without rigorous justification, at least with respect to use of ED for assembly Consolidation of many contiguous but separately owned parcels of land under one owner supposedly creates a holdout problem, with each seller having an incentive to hold out to be the last to settle and capture any rent accruing to the assembly Because of either monopoly prices of sellers or high transactions costs or both (since the buyer can trade off between the two), the free market results in a suboptimal amount of assembly being undertaken The next step in the argument, although not in logic, is that ED improves the situation

A crude examination of the circumstances in which ED is actually used does not leave the impression that their outstanding common characteristic is consolidation of ownership rights, which is nowhere mentioned as a necessary condition of the granting of ED power In practice, almost all departments of federal, state, and local government, many regulated industries, and government-related educational and medical establishments have some form of ED power in most states, regardless of whether they wish to acquire one parcel for a schoolhouse or campaign headquarters or 1,000 miles of right-of-way for a freeway The outstanding characteristic of situations where ED is used is land acquisition by a government-related body, not consolidation of ownership rights [5] The two sets, although far from mutually exclusive, are not perfectly overlapping Notable examples of industries in the United States which must assemble property, do serve the public, but have not themselves been granted ED are agriculture, private manufacturing, and suburban development [6]

---

[2] To contest the taking, as opposed to the prices paid, the condemnee must prove that the condemnor has used his power "unconstitutionally," "fraudulently," or "abusively " The courts may, however, review the original grant of ED on the grounds that it is not for a legitimate public use

[3] "Transaction costs" here include court costs It is questionable whether the first precondition of the Coase theorem, exclusively assigned property rights, is met in most ED contexts, since the condemnor is usually a government-associated agency

[4] E g , Arrow (1970), Colean (1970), and McCloskey (1972)

However, since the economic efficiency argument for ED rests on its comparative advantage over the free market in assembling separately owned parcels into a single ownership unit, this is the focus of the present study A necessary condition for ED to be an efficient institution is that the welfare costs, due to deviations between price paid under ED and the value of a parcel in its best alternative use, plus resource costs of transacting, be less than under alternative property right assignments After a brief theoretical comparison of the relative efficiencies of ED and the free market, a model of price determination under ED is described and then tested with data from urban renewal in Chicago A comparable set of data from free-market assemblies was unobtainable However, theoretical considerations and the limited evidence available offer no support for the asserted efficiency of ED

Theoretical Comparison of ED with Assembly in the Free Market

Consider an area where there are many homogeneous properties Assuming the market functions efficiently in transferring properties to their highest-valued uses, subject to imperfect information and positive search costs, the atomistic market reservation prices[7] of current owners will form a distribution skewed to the right. Transactions will take place within a range around a mean market value, which will be below the modal value of the distribution. Then, for purchases of randomly scattered parcels with an average amount of search per parcel, the supply curve is infinitely elastic at a price equal to mean market value. The supply curve of contiguous properties to any buyer, however, is an increasing function of the number of parcels to be purchased in a given period, because of the rising probability of encountering owners not at the selling margin whose reservation price exceeds mean market value of the property.[8] The supply curve of parcels for assembly is steeper the greater the dispersion of atomistic reservation prices of current owners in the market.

---

[5] A distinction should perhaps be made here between grants of ED for which government relatedness seems to be a necessary condition and actual exercise of the latent power by use or threat of use of condemnation proceedings It is possible that ED is more frequently exercised by bodies which have the latent power when assembly of many parcels is involved I know of no body of data to test this hypothesis Two pieces of crude evidence tend to support the presumption that ED is most valuable in an assembly context (1) Although most government-related bodies now have the power, the earliest to acquire it were those engaged in assembly, e g , gristmills, canals, roads, and railroads (2) In terms of area taken, the most extensive current uses are for highways, urban renewal, and the U S Army Corps of Engineers, all of which typically involve assembly (U S Congress 1965)

[6] These industries, however, have presumably benefited from particular uses of ED e g , agriculture, from the use of ED for drainage, gristmills, and railroads, and private manufacturing, from the use of ED for highways Similarly, although only those urban renewal projects sponsored by a public agency are entitled to use ED, it might be more correct to identify the "user" of ED as the private real estate industry, to which land acquired under ED by the urban renewal authorities is ultimately sold at less than the acquisition cost. In other words, some public agencies may be, more or less, fronts created by private industries to accommodate the constitutional "public use" requirement Thus, it may be misleading to identify ED with government-related agencies

Figure 1 represents supply and demand conditions of a typical buyer for parcels for one type of project within the universe of projects requiring assembly, for example, urban renewal or highways. *MCm* is the supply curve of randomly scattered parcels, and *MCa* is the supply curve of contiguous properties. Both reflect atomistic reservation prices and social opportunity costs. The slower the rate of acquisition, the flatter *MCa*, approaching *MCm* in the limit. The demand curve is the marginal value product curve per parcel of a given size, as part of an assembly, for a given rate of acquisition and set of initial conditions with respect to number of buyers in the industry. The slower the rate of acquisition, the lower the demand curve.[9]

If the buyer cannot discriminate between sellers (because reservation prices are not freely observable), and so expects to pay all *n* sellers the reservation price of the *n*th, *MCa* is the buyer's average cost curve. Profit

---

[7] "*Atomistic market reservation price*" is defined as the reservation price of a seller to a buyer whom he did not suspect of planning an assembly.

[8] Assume that assembly involves drawing a sample of size *n* from the distribution of atomistic reservation prices. Let Y_i denote the *i*th order statistic of this sample, i.e., $Y_1 < Y_2 < \cdots < Y_n$. The height of the *MCa* curve, the supply curve of contiguous parcels, at *n* parcels is the expected value of the *n*th-order statistic from a sample of size *n*. For any point *i* between the origin and the point *n*, the height of *MCa* measures the expected value of Y_i. Strictly, there is a different *MCa* curve for each sample size. The precise shape of the curve is not important. It is sufficient to establish that it is upward sloping. This follows from the fact that for many density functions, e.g., normal, gamma, and exponential, the expected value of the *n*th order statistic has the following form

$$E(Y_n) = \theta + \beta(n)\sigma,$$

where θ = mean of underlying distribution, σ = standard deviation of underlying distribution, and $\beta(n) > 0$, $\beta' > 0$, $\beta'' < 0$. The exact form of $\beta(n)$ depends on the specific distribution.

[9] If the project has some minimum feasible size, the demand curve will be discontinuous at the corresponding number of parcels. If assembly of dispersed ownerships into one unit permits the internalization of externalities, there may be increasing returns to scale, implying an upward-sloping demand schedule over an initial range. The shape of the demand curve may affect the conclusion but not the method of analysis of the respective welfare costs of the free market and ED.

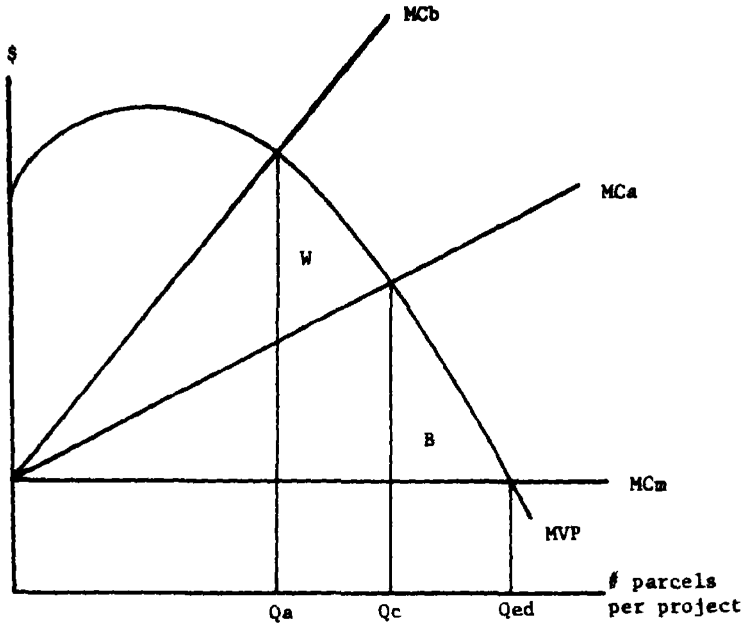


FIG 1 —Welfare costs of assembly with and without eminent domain

maximization implies *Qa* parcels will be bought A resource misallocation cost equal to the triangle *W* is implied Rents accrue to the buyer and to intramarginal sellers

In the absence of barriers to entry by other buyers and sellers on alternative sites, these quasi-monopsonist and scarce-factor rents cannot persist in the long run Competition among buyers, attempting to capture the rent which accrues to intramarginal sellers in the absence of discrimination, will lead to the development of techniques to discover true seller reservation prices This is facilitated by competition among sellers on alternative sites The incentive to conceal atomistic reservation price exists only if substitutes are not available at competitive prices

With competition on both sides of the market, seller reservation prices to an assembler will be no different from atomistic reservation prices This implies that *MCa* is the effective marginal cost curve The profit-maximizing number of parcels per project is *Qc* This is consistent with the criteria of social optimality, although it is less than *Qed* at the intersection of the *MVP* curve with the market supply curve of isolated parcels, because the social opportunity cost of contiguous properties exceeds the social opportunity cost of the same number of randomly scattered properties

In the absence of perfect substitutes for a particular site, a rent exists that is potentially capturable by the owners of land. However, a holdout problem, as distinct from a monopoly problem, is implied only if the rights to the rent are not exclusively assigned For example, if uniqueness is obviously specific to a particular parcel, the rent will accrue to its owner, and there is no reason for the owners of other inputs to raise reservation prices above atomistic market level. In general, there may be less than perfect substitutes for many or all parcels in a group, that is, for the ith of n parcels in a group, A, there exist alternative parcels which may be substituted for it at a cost to the assembler x_i percent higher, where x_i may vary across the n parcels If each seller in set A anticipates that sellers of the alternative parcels will settle at their atomistic reservation prices and that other sellers in set A will settle for not less than their share of the rent as defined by substitution possibilities, prices and shares of the rent are determinate Assembly is characterized by monopoly prices (or rents to superior factors), but it is misleading to speak of a holdout rather than a monopoly problem, since changing the unit of ownership is immaterial

A distinct holdout problem, specific to assembly, arises only if at least one seller anticipates that others will settle for less than their "share" of the rent and therefore tries to capture more than his "share," up to a maximum for any one seller of the entire rent net of competitive prices for other inputs Moreover, the existence of a potential rent, due to lack of perfect substitutes, may occur in assembly even though there are physically equivalent sites for a particular project For competition to be effective, substitutes must be available at competitive prices throughout the negotiation process But assembling many ownerships is typically a sequential process, because transacting is time consuming When some parcels on site A have already been acquired, in negotiations for the outstanding parcels, site B, assumed physically equivalent, is a less than perfect substitute unless the expected cost of B gross of transactions costs is no greater than the expected cost of the remainder of A gross of transactions costs plus the resale value of parcels already acquired on A net of transactions cost Thus, the costs of negotiating many contracts and of uncertainty as to prices, rather than physical availability, may limit the operation of competitive forces in an assembly context The implied potential rent to outstanding parcels on a particular site when some have already been acquired creates an incentive to hold out [10]

In conclusion, a problem which is specific to the assembly of several ownerships into a single unit arises as a combination of a monopoly and free-rider problem If a particular site is physically unique, rents will accrue to current owners as to any scarce factors Even in the absence of physical uniqueness, assembly may create a potential rent to the last parcels to settle, due to transactions costs and uncertainty as to prices of alternatives. Even if property rights are sufficiently well defined for the market to function efficiently when changing the unit of ownership is not in question, the rights to any rents generated by changing the unit of ownership may not be well assigned, since reservation prices are not freely observable and there may be uncertainty as to prices charged by other sellers This creates a type of free-rider problem, each party having an incentive to try to capture more than his atomistic reservation price for his input, on the assumption of asymmetrical behavior by other parties.

---

[10] *The options contract is a method of maintaining competitive pressures throughout the negotiation process*

Where rents are attributable to physical uniqueness, they are no different from rents to scarce factors No resource misallocation is implied, but assembly may entail a welfare cost not generally associated with scarce factors if the rent is dissipated in bargaining over its distribution However, where potential rents are attributable solely to the transactions costs of maintaining competitive pressures, assembly entails resource misallocation and welfare costs similar to those implied by the monopsony solution in figure 1 If holdout behavior is anticipated, *MCa*, reflecting atomistic reservation prices, is no longer the relevant marginal cost curve Expenditure on devices to circumvent or eliminate the incentive to hold out will be incurred Such devices include concealment of the identity of the buyer, the purpose and extent of the planned assembly and prices paid for parcels, and the use of brokers and special contractual forms, such as options or uniform price contracts The optimal level of such expenditure defines a new *MC* curve, lying above *MCa* but not necessarily equal to *MCb*, and a new free-market level of output and welfare cost, which may exceed or fall short of *W* in figure 1 [11] The expenditure on transactions to avoid the holdout problem represents an additional waste due to assembly, if efficiency is defined relative to the zero transactions cost situation

If the free market tends to produce suboptimal-sized assemblies, the *ED* "just compensation" standard may have the opposite effect Just compensation is defined as "fair market value" exclusive of the value to this particular seller If enforced, this implies prices equal to *MCm* in figure 1 If the demand curve is assumed unchanged, optimal output, at *Qed*, is excessive Fixing the price at the market average understates the social opportunity cost The welfare cost, given by the triangle *B*, may be less than or exceed *W*

Note that the welfare costs of ED are likely to be high in precisely those circumstances in which the market is inefficient, that is, where seller reservation prices are dispersed, implying an inelastic social opportunity cost curve, *MCa*, and high transactions costs of discriminating between sellers. A similar symmetry exists in the absence of transactions costs. In the free market, the buyer would discriminate along the *MCa* curve. Under ED, if the expected court award, *MCm*, were less than *MCa*, the value of the parcel to the seller, he would offer the buyer the difference to prevent the taking of the parcel. Then *MCa* would be the effective marginal cost curve to the buyer, and the profit-maximizing output level would also be socially optimal. Thus, with zero transactions costs, both the free market and the ED assignments of property rights yield an optimal result. Since transacting under both regimes is not costless, the efficiency conclusion on ED relative to the market is theoretically ambiguous, even if fair market value is paid.

---

[11] The various devices developed in the free market to circumvent the holdout problem are discussed in Munch (1973)

Price Determination under ED

Constitutional dictates notwithstanding, since fair market value is not a freely observable datum, both buyer and seller will invest in searching for the most favorable price, and the outcome may differ systematically from fair market value, depending on the costs of and returns to search for the two parties. For both, the optimum search strategy implies setting a reservation price for settlement out of court which maximizes expected wealth, net of transactions costs, over the two alternatives, settlement in or out of court.

Price determination under ED may be formulated by the following reduced-form model:

$$S^b = \bar{S}^b, \tag{1}$$

$$S^s = \bar{S}^s, \tag{2}$$

$$C^b = C^b(P_m), \tag{3}$$

$$C^s = E[k(P_c - P_o)], \tag{4}$$

$$P_c = P_m - h^b(C^b) + h^s(C^s) + v;\ v \sim [0, \sigma_v^2(P_m)], \tag{5}$$

$$E(P_c)^b = P_c + w,\ w \sim [0, \sigma_w^2(P_m)], \tag{6}$$

$$E(P_c)^s = P_c + z,\ z \sim [0, \sigma_z^2(P_m)], \tag{7}$$

$$P_o^{\max} = E(P_c)^b + (C^b - S^b), \tag{8}$$

$$P_a^{\min} = E(P_c)^s - (C^s - S^s), \tag{9}$$

$$P_o^{\max} - P_a^{\min} = E(P_c)^b - E(P_c)^s + C^b + C^s - S^b - S^s \geq 0, \tag{10}$$

$$P_v = \tfrac{1}{2}(P_o^{\max} + P_a^{\min}) \tag{11}$$

$$= \tfrac{1}{2}[E(P_c)^b + E(P_c)^s] + \tfrac{1}{2}[(C^b - S^b) - (C^s - S^s)]$$

$$= P_c + \tfrac{1}{2}[\Delta C^b - \Delta C^s], \tag{11'}$$

where S^b = buyer's costs of out-of-court settlement, C^b = buyer's court costs, S^s = seller's costs of out-of-court settlement, C^s = seller's court costs, $\Delta C^b = C^b - S^b$, $\Delta C^s = C^s - S^s$, P_c = price awarded in court, $E(P_c)^b$ = buyer's expectation of court award, $E(P_c)^s$ = seller's expectation of court award, P_m = market value, P_o^{max} = buyer's maximum offer in out-of-court negotiations, P_a^{min} = seller's minimum ask price in out-of-court negotiations, P_v = price determined voluntarily in out-of-court settlement, and v, w, z = stochastic error terms All variables are measured relative to market value

Equations (1) and (2) state that settlement costs are a constant proportion of market value Equation (3) makes buyer court costs a variable function of market value Equation (4) states that seller court costs are the expected value of the attorney's fee, which, in a contingent fee contract typically used in ED cases, is some fraction of the difference between the court award and the buyer's final offer

Equation (5) describes the determination of price in court Any systematic deviation of court award from market value is the result of expenditure by both parties on court inputs and the function $h^b(\ )$ and $h^s(\ )$, which relate this expenditure to influence on court verdict Equations (6) and (7) state that both parties have unbiased expectations of court award, with random errors having zero mean and variance a function of parcel value The assumption that expectations are unbiased is plausible if both hire legal counsel with experience in the field of condemnation

Equations (8) and (9) state the upper and lower bounds, respectively, on offer and ask prices consistent with wealth maximization[12] Equation (10) is a necessary condition for settlement out of court, that is, $P_o^{max} \geq P_a^{min}$ Rearranging terms produces

$$(C^b - S^b) + (C^s - S^s) \geq E(P_c)^s - E(P_c)^b = z - w$$

Thus, settlement out of court requires that the sum of the incremental costs due to going to court exceed the difference between the seller's and buyer's expectations of court award Note that this is independent of the assignment of liability for court costs

Equation (11) describes the determination of price in an out-of-court settlement It makes the simplest assumption that the parties split the difference between P_o^{max} and P_a^{min} The subsequent analysis applies if $\frac{1}{2}$ is replaced by any positive fraction, to reflect "unequal bargaining power," and the Δ in equation (11') omitted, since settlement costs are assumed to be an invariant fraction of market value

---

[12] Actual final offers and asks may lie within these bounds, depending on each party's assessment of the probability of going to court

The Alternative Hypotheses

Given this model, one set of sufficient conditions for prices paid under ED to approximate market value with random errors is the following: (i) Anticipation of ED does not affect market value prior to the filing of the petition to condemn It is fair market value as of the date of filing the petition to condemn that the courts are instructed to award The timing of filing is at the discretion of the condemnor (ii) $C^b(P_m) = C^s(P_m)$, that is, buyer and seller court costs bear the same functional relation to market value (iii) $S^b = S^s$ (iv) $h^b(\ ) = h^s(\ )$, that is, the returns to court expenditure are the same for buyer and seller

These conditions yield the null hypothesis that prices paid under ED approximate market value with random errors, regardless of whether the price is negotiated voluntarily or awarded in court An alternative hypothesis may be generated by incorporating into the model the following assumptions based on economic theory or empirical evidence from urban renewal in Chicago (i′) Anticipation of ED tends to depress market value ED implies a loss of use rights associated with the property, in particular, effective loss of the right to refuse to sell and greater uncertainty of lease duration, since condemnors have special rights to terminate leases Thus, anticipation of ED would tend to depress market value, *ceteris paribus*[13] (ii′) $C^b(P_m) \neq C^s(P_m)$, that is, optimum expenditures on court costs are not the same for buyer and seller

Optimum expenditure on court costs depends on the costs of and returns to hiring legal "inputs" and constraints on choice of quality and quantity The buyer in condemnation cases is typically constrained by statutory requirements on the use of legal personnel Thus, the Department of Urban Renewal in Chicago is represented in all court cases by the City of Chicago's corporation counsel and is required to obtain at least two independent appraisals in each case. The seller, on the other

---

[13] If at time t_0, condemnation is anticipated at some future date, t_n, the present value of the property at t_0 is

$$PV_{t_0} = \frac{E(Pt_n)}{(1 + r)^n} + R\,\frac{1 - (1 + r)^{-n}}{r},$$

where $E(Pt_n)$ = expected price obtainable at t_n under ED and R = value of service stream derived from property from t_0 to t_n Uncertainty as to the value placed by the courts on maintenance expenditure may reduce the optimum maintenance program and reduce R, if the property is rented, this will be exacerbated by a reduction in the demand for leases because of uncertainty as to duration Thus, even if the courts award the market value of the property, measured as the present value of its expected income stream at t_n when condemnation is filed, this PV_{t_n} will be lower than it would have been in the absence of anticipated condemnation This may be offset by defensive expenditures on "improvements" that are valued by appraisers and the courts higher than they would have been by the market, or by arranging dummy sales at inflated prices to be used in court as evidence of market value Thus, the net effect of anticipation of ED on property value is ambiguous a priori However, since the properties acquired for urban renewal in Chicago are predominantly rented, the former negative effect is expected to dominate hand, presumably adjusts the quality (or implicit hourly wage rate)[14] of attorneys and witnesses commensurate with the stakes of the case

However, on the other dimension of court expenditure—hours per case—it seems likely that the seller faces greater constraints than the buyer. Handling any case requires some minimum number of hours of appearance in court independent of the value of the property being litigated This number of hours times the implicit hourly wage rate of the lowest-quality lawyer represents a minimum fixed cost of going to court for the seller The buyer, on the other hand, may have greater flexibility in reducing hours per case because of the possibility of grouping several similar parcels in one case The possibility of spreading the fixed cost of a court case over several parcels is available on equal terms to sellers only if the costs of negotiating to hire the same lawyer are zero

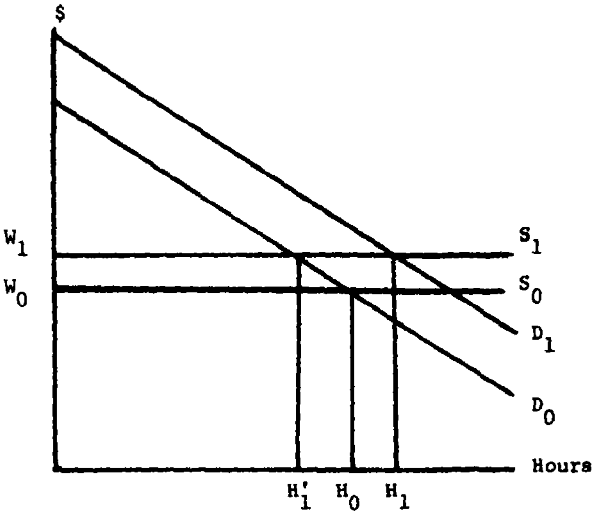
If we ignore for the moment constraints on hours per case, constraints on the buyer's choice of quality of lawyer will have the effect of raising the buyer's expenditure on low-valued properties and lowering it on high-valued properties relative to the seller's expenditure The effect is illustrated for low-valued properties in figure 2

Figure 2 represents the costs and returns to hours per case for the buyer and seller The null-hypothesis supply and demand curves, S_0 and D_0, are drawn on the assumptions that both parties face unconstrained choices, that both correctly anticipate the other's behavior and adjust accordingly, and that courts are neutral Both would have identical expenditures, W_0H_0

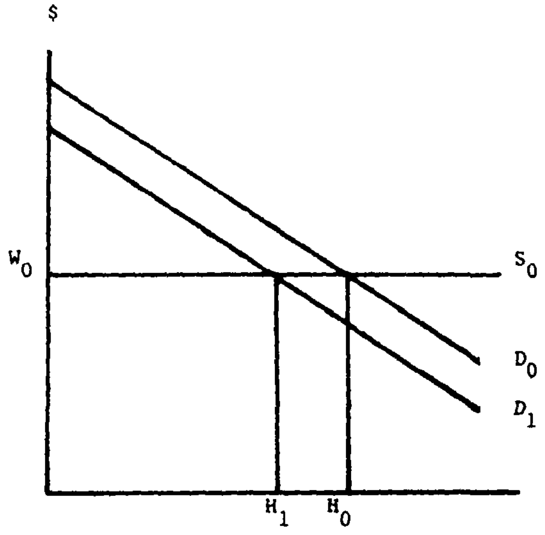
The curves S_1 and D_1 are drawn on the assumption that the buyer is constrained to employ above-optimum-quality lawyers at a higher wage rate, W_1 Assuming the seller's choice of quality unchanged at W_0, the buyer's lawyers will be relatively more effective, implying an upward shift in the buyer's returns-per-hour curve and a downward shift in the seller's returns-per-hour curve to D_1 Thus, the effect of the constraint is to raise buyer expenditure on court influence above that of the seller ($W_1H_1 > W'_0H_1$) Given the buyer's constrained choice, it may be optimal for the seller to raise quality above W_0 But this need not result in an equal level of expenditure If quality levels were equalized at W'_1, the returns curves would shift back to D_0, the curves representing returns to equal-quality hours Optimal-hours input would be H_1 Thus, the same court outcome would be achieved with a more quality intensive input mix But if hours per case are not flexible downward but are constrained at some minimum above H'_1, the higher-quality lawyer would require a higher fraction of the award, implying a downward shift in the seller's demand curve, and reduced expenditure

---

[14] Although on the seller's side the form of contract is typically a contingent fee or share contract, not a fixed hourly wage contract, the seller can vary the implicit hourly wage rate and hours per case indirectly by choice of attorney and by varying the attorney's share of the final award



Buyer



Seller

FIG 2 —Expenditure on court inputs

Thus, constraints on reduction in quality of legal inputs by the buyer and on hours per case for the seller will tend to result in the buyer's expenditure on court costs exceeding the seller's on low-valued properties. For high-valued properties, the quality ceiling produces the reverse effect. The hours constraint is inoperative.

A third force tending to produce the same result is the existence of economies of scale for the buyer. Court recognition of precedent implies an outward shift in the returns-per-hour curve for parcels which are sufficiently homogeneous for a precedent effect to operate. In the urban renewal sample, homogeneity, and hence the buyer's relative advantage due to precedent economies of handling many parcels, is expected to be greater on low- than on high-valued parcels. Again, transactions costs are assumed to prevent sellers from enjoying these scale economies to the same degree

These assumptions about the structure of court costs and returns imply:

$$C^b > C^s \text{ on low-valued properties,}$$

$$\frac{\partial C^b}{\partial P_m} < \frac{\partial C^s}{\partial P_m} \leq 0$$

Incorporating into equations (5) and (11′) yields the alternative hypotheses

1) $$P_c < P_m \text{ on low-valued properties,}$$

$$\frac{dP_c}{dP_m} = \frac{-d[k^b(C^b)]}{dP_m} + \frac{d[k^s(C^s)]}{dP_m} > 0$$

Thus, in the absence of any court bias, high-valued parcels will tend to receive a higher fraction of market value in court than will low-valued parcels

2) $$\frac{dP_v}{dP_m} - \frac{dP_c}{dP_m} = \frac{1}{2}\left(\frac{dC^b}{dP_m} - \frac{dC^s}{dP_m}\right) < 0$$

Prices reached voluntarily in out-of-court settlements rise less, relative to market value, than do court awards

Empirical Evidence

The empirical estimates of the relationship between ED prices and market value are based on land acquisitions by the Chicago Department of Urban Renewal for three large projects during the period 1962–70[15] The data obtained from HUD consist of final price paid (*Ped*), assessed value for tax purposes (*TAVL*), and date of acquisition (*DATE*) for all parcels. In addition, for all parcels in two of the three projects, method of acquisition —voluntary settlement or court—is known In order to compare price actually paid with market value, an estimate of market value was derived for each parcel in the ED sample The procedure used was to estimate a

---

[15] The projects are Southeast Englewood, Near West Side, and Lincoln Park I.

TABLE 1*

MARKET SAMPLE SUBDIVIDED BY ZONE, REGRESSIONS OF *PRL*

Zone	C	$TAVL$	$DATE$	TEN	$VALL$	DC	DCT	DZ	DZT	R^2	SEE
$R2$ and 3 r = 674(1)	7.2348 (50.16)	0.3109 (17.005)								3008	2228
$Z = R3$ n = 637(1′)	6.2324 (9.508)	0.32304 (4.182)	0.00396 (5.517)	0.03495 (3.186)	0.17033 (3.526)	0.04878 (0.597)		0.33513 (0.5270)	−0.03805 (−0.4789)	3648	2133
$R4$ and 5 n = 423(2)	7.35623 (33.9)	0.28457 (10.916)								2206	42557
$Z = R5$ n = 72(2′)	5.31613 (10.79)	0.22607 (7.465)	0.00678 (4.165)	0.03419 (3.074)	0.49279 (5.399)	0.03213 (0.23008)		0.16596 (0.3011)	−0.00905 (−0.1344)	3354	.3958
$B1$ and 2 n = 64(3)	7.57294 (18.5)	0.26047 (5.566)								.3332	3587
$Z = B2$ n = 55(3′)	7.80816 (6.36)	0.05664 (0.412)	0.00462 (1.326)	0.00109 (0.03668)	0.27981 (2.4468)	−1.59456 (−1.482)	0.156829 (1.347)	−1.47020 (−1.255)	0.173948 (1.26021)	4542	344567
$C1$ and 2 n = 30(4)	8.94112 (15.2)	0.086657 (1.292)								0562	.40077
$Z = C2$ n = 5(4′)	7.54587 (5.37)	0.103007 (1.011)	0.011168 (1.8275)	−0.10787 (−1.0255)	0.251795 (1.1031)	−0.916476 (−1.656)	0.082346 (2.005)	−6.35926 (−1.5718)	0.714092 (1.5336)	4960	.338189
$R4$ n = 351	5.41779 (12.6380)	0.26699 (9.4933)	0.007903 (5.2367)	0.0313 (1.6916)	0.402177 (4.8588)					3757	.330719
$R5$ n = 72	4.35524 (2.5109)	0.11731 (1.48779)		0.051675 (2.3779)	0.88131 (2.53988)					.2555	.624102
B n = 64	6.57099 (11.1487)	0.22758 (4.8762)	0.00535 (1.61136)		0.2320 (2.11887)					.4112	.34266
C n = 30	6.75036 (5.75987)	0.156731 (1.79205)	0.01059 (1.73873)		0.3066 (1.3799)	−1.11580 (−2.15098)	0.08720 (2.13507)			.4707	.33858

* All variables defined in table 1A.

TABLE 2

ED SAMPLE SUBDIVIDED BY ZONE
(t-STATISTIC IN PARENTHESES)

ZONE	A REGRESSIONS OF Ped			B REGRESSIONS OF $\hat{P}m$		
	C	$\hat{P}m$	R^2/SEE*	C	Ped	R^2/SEE*
$R4$ n = 202	−11 2807 (−10 4540)	2 20652 (19 0102)	6437 (430314)	6 60185 (46 493)	0 291745 (19 01)	6437 (0 15647)
$R5$ n = 192	−2 71576 (−1 28492)	1 27976 (5 75121)	1476 (696318)	8 40721 (44 2836)	0 115344 (5 75124)	1476 (0 209045)
C n = 232	−6 01600 (−3 1975)	1 64972 (8 43511)	2363 (70775)	8 20651 (48 909)	0 143216 (8 43517)	.2363 (2 0853)
B n = 147	−9 5534 (−7 24648)	1 99435 (14 7574)	6003 (48559)	6 77291 (33 465)	0 301006 (14 757)	6003 (0 188652)

	$\hat{P}m_{min}$	$\hat{P}ed_{min}$	$\hat{P}m_{mean}$	$\hat{P}ed_{mean}$	$\hat{P}m_{max}$	$\hat{P}ed_{max}$
$R4$	6,180	2,920	11,272	9,200	35,260	135,000
$R5$	6,464	4,985	13,619	13,000	22,100	24,200
C	7,729	6,340	15,449	20,100	28,300	53,700
B	9,045	5,515	17,951	22,100	38,600	98,800

NOTE —*Definition of variables*

$\hat{P}m$ (min), $\hat{P}m$ (mean), $\hat{P}m$(max) = minimum, mean, maximum estimated market values from $\hat{P}m = Xed\hat{\beta}$

$\hat{P}ed$(min), $\hat{P}ed$ (mean), $\hat{P}ed$(max) = minimum, mean, maximum predicted ED prices given $\hat{P}m$ and regression coefficients estimated from regression of Ped on $\hat{P}m$ (first set above)

* SEE = standard error of estimate

relationship between market value, assessed value, and several property characteristics from a sample of properties sold on the free market in areas of Chicago similar in character to those where urban renewal projects have been undertaken[16] The estimating equations used are the last four reported in table 1 The coefficients obtained were then applied to data on the same explanatory variables to generate an estimate of market value, $\hat{P}_m$, for all parcels in the ED sample

Table 2 reports the results of ordinary least-squares regressions of price paid under ED on estimated market value Since the equation is estimated in logarithmic form, the null hypothesis predicts a zero intercept and coefficient of unity on $\hat{P}_m$ The evidence strongly supports the alternative hypothesis, that low-valued properties receive less than market value and high-valued properties receive more than market value.[17]

---

[16] The market sample consists of property sales in Chicago over the period 1968–72 in the Community Areas (as defined by the 1960 census) in which some urban renewal project has been located Sample size considerations dictated drawing from a much larger area than that covered by the ED sample, which includes only three of the 20 or more urban renewal sites Theoretically, it is appropriate to draw the market sample from areas similar in character to the ED sample but not sufficiently close to actual urban renewal sites for property values to have been affected, since ED compensation is not supposed to reflect value of the property to the buyer In practice it was found that private market sales in the immediate vicinity of urban renewal sites were very sparse, so contamination of the market sample by the effects of urban renewal should be minimal

To test the plausibility of the results and obtain an idea of the order of magnitude involved, the mean, minimum, and maximum $\hat{P}_m$ were calculated for each zoning class and compared with the corresponding values implied by the regression equations for ED prices. The exact results vary across zones, but the generalization that ED is a tax on low-valued and a subsidy on high-valued properties holds for all zones. As a rough approximation, a \$7,000 parcel receives about \$5,000, a \$13,000 property breaks even, and a \$40,000 property may get two or three times its market value.

The regressions in table 3 are designed to investigate which property characteristics are associated with deviations of ED price from market value ($PDIF = Ped - \hat{P}m$). If deviations of ED prices from market value were due solely to factors common to all properties of a particular value, the coefficients in the *PDIF* equations should be insignificantly different from zero. The consistently positive coefficients on *TAVL* are surprising *if* $\hat{P}_m$ *accurately incorporates all of the systematic relation between* tax-assessed value and market value. A possible explanation is that either appraisers or the courts treat assessed value as an indicator of market value and fail to recognize that component of actual assessed value which is a deviation from the average market assessment ratio and which would be capitalized into property value in a free-market sale because of the effect on tax liability. This would imply a coefficient closer to unity for the ED than for the market sample, which is in fact found.[18] While there may be some validity to this explanation, it is insufficient alone to account for the evidence on deviations of ED prices from market value. If courts

---

[17] This estimate is no doubt affected by, but cannot be readily explained away in terms of, either random or systematic error in $\hat{P}_m$. The purpose of using a predictive equation for $\hat{P}_m$ is to control for any systematic variation across properties in the ratio of assessed value to market value, and hence avoid a systematic bias in the estimate of market value that would exist if assessed value alone were used as a simple proxy for market value. Systematic bias in the predicted $\hat{P}_m$ may nonetheless exist, since measurement error in the explanatory variables of the predictive equation will lead to biased estimates of the coefficient vector used to predict $\hat{P}_m$. The direction of the bias is not known, however, in the case of a multivariate ordinary least squares regression with errors in more than one independent variable. Assuming that the errors in $\hat{P}_m$ are additive in the logarithmic specification and uncorrelated with either the true values or the stochastic disturbance term in the relation between ED price and true market value, then a lower bound on the true value of the slope coefficient is given by the estimate obtained by regressing P_{ed} on $\hat{P}_m$ (set A in table 2) and an upper bound by the reciprocal of the coefficient obtained by regressing $\hat{P}_m$ on P_{ed} (set B in table 2). Both exceed unity. However, if the assumptions with respect to the error structure in $\hat{P}_m$ are not met, this test does not yield bounds on the true coefficient. In the absence of knowledge of the direction of bias in $\hat{P}_m$, nothing can be said a priori about the direction of potential bias in the estimated relation between ED price and market value.

[18] The total coefficient on *TAVL* is the sum of the coefficient of the *PDIF* equations (table 3) and that from the market equations (table 1) used in calculating $\hat{P}_m$.

TABLE 3

ED SAMPLE REGRESSIONS OF $PDIF = Ped - \hat{P}m$

Zone	C	$TAVL$	DC	DCT	TEN	$DATE$	$VALL$	R^2	SEE	D-W
$R4$ $n = 203$	−3 74440 (−12 2364)	0 457536 (12 069)						4202	40588	1.446
	−2 49465 (−2 7131)	0 438037 (8 6888)	1 09439 (0 9387)	−0 13135 (−1 0111)	0 054508 (2 6284)	−0 11790 (−1 787)	−0 26440 (−1 1276)	4613	396182	1 5377
$R5$ $n = 192$	−5 29391 (−21 135)	0 66307 (21 02)						6993	38425	1 4726
	−3 58377 (−5 373)	0 654012 (17 2521)	0 829029 (0 9892)	−0 115636 (−1 1976)	0 02035 (0 98535)	0 005178 (2 96548)	−0 365624 (−3 26795)	7315	36797	1 4762
C $n = 232$	−4 92328 (−16 4360)	0 59878 (17 41)						5686	46737	1.6456
	−4 13779 (−8 2683)	0 51725 (10 9138)	2 2541 (3 607)	−0 20778 (−2 9596)	0 06961 (4 9107)	−0 00895 (−2 7589)	−0 04547 (−0 5782)	6941	3979	1.8546
B $n = 147$	−3 75475 (−9 7705)	0 44967 (10 215)						4185	43675	1.965
	−3 81760 (−7 526)	0 498201 (6 954)	1 3078 (1 51186)	−0 14715 (−1 5153)	0 098495 (3 59549)	0 00477169 (1 54103)	−0 107667 (−1 37942)	5087	408545	2 0402

were guided solely by assessed value, there would be no reason for sellers to engage lawyers and appraisers to defend them in court Moreover, the adherence-to-assessed-value hypothesis cannot account for the divergence between the pattern of court awards and settlement prices in relation to market value predicted by the court-costs theory and confirmed by the evidence discussed below.

Other variables of interest are *TEN*, the number of tenants with listed phones, and *DATE*, the date of acquisition, measured in months from earliest observation in each location

In the market sample, *TEN*, the number of tenants with listed phones, is expected to measure a combination of tenant capacity and current occupancy If tenant capacity and potential income vary among structures to an extent not fully captured in assessed value, the positive coefficients observed in the predictive equation for the two residential zones are to be expected The significant positive coefficients on *TEN* in three of the four *PDIF* equations then suggest that high capacity or occupancy rates have a greater impact on ED awards than on free-market prices If it is the case that anticipation of ED reduces occupancy rates due to uncertainty of lease duration and compensability of improvement and maintenance expenditure and that this is reflected in lower ED awards, then ED creates an incentive for the buyer to delay acquisition or hold out to depress prices analogous to the seller's supposed incentive to hold out in the free market [19]

However, if the effect of ED, whether intended or not, is to reduce occupancy rates below normal, then the observed number of tenants will measure with error the potential rental capacity of properties in the ED sample Then the positive effect of rental capacity on market value will be biased down in the estimates of market value If courts respond to potential rather than actual occupancy, a positive coefficient on *TEN* in the *PDIF* equations may be expected Given this potential for errors-in-variables bias if the effect to be measured in fact exists, it seems, in principle, not possible to measure the holdout incentives of buyers to depress occupancy and hence prices under ED in the absence of better data on potential rental capacity

---

[19] Residential buildings are predominantly tenant occupied in urban renewal areas of Chicago, with average period of occupancy 1–3 years Anticipation of the effect of ED by sellers presumably includes awareness that high vacancy rates may depress ED prices This would create an incentive to maintain occupancy rates by lowering rental charges The extent to which such defensive behavior by sellers is a good investment depends on the relative magnitudes of the price elasticity of demand for leases, the reduction in demand for leases due to uncertainty as to duration, and the elasticity of ED prices with respect to number of tenants and rental rates An upper bound on the estimated effect on ED price of an increase by one in the number of tenants lies between 5 and 10 percent As a rough calculation, at the point of means for the court sample (table 5), an increase in the number of listed tenants from one to two would raise ED price by $600–$1,200 For any individual landlord, the demand for leases is likely to be inelastic, since he cannot assume other landlords will maintain higher rentals as he reduces his if anticipation of ED is widespread in the neighborhood The less elastic the demand curve, the greater the reduction in rental rates required to increase the number of tenants, and the greater the loss in rental income over the interim between anticipation of ED and filing the petition to condemn Thus, stocking up with tenants in anticipation of ED is only likely to be profitable if the demand curve is elastic, or, if inelastic, where the expected delay is short, so that loss of rental income is less than expected gain in ED price

The coefficients on the *DATE* variable only support the weak conclusion that there is no evidence that ED holds down the rise in prices over the duration of the assembly. The coefficients must be interpreted as the difference between the monthly rate of inflation in the market and ED samples, since the former was incorporated into $\hat{P}_m$ (except in *R*5, since for this zone *DATE* was insignificant in the market sample) But the market rate of increase of property values differs across the city at any point in time and over time The average market rate of inflation was probably less in the early 1960s, from which roughly half the ED sample was drawn, than in 1968–72, from which the market data were drawn Therefore, ignoring interarea differences, the null hypothesis that ED prices follow the same pattern over time as the market would imply a negative coefficient on *DATE* in the equations for *PDIF* The size should be the difference between the monthly market rate of inflation in 1968–72 and 1962–65 weighted by the proportion of the ED sample drawn from the earlier period Taking 5 percent and 2 5 percent as the average annual rates for the two periods, weighting the difference bv 0 5 yields a rough estimate for the expected coefficient on *DATE* of −0 001 Of the estimated coefficients, two exceed and two fall short of this benchmark for the null hypothesis To the extent prices under an assembly in the free market would rise faster over time than the market average due to the holdout problem, the conclusions are more favorable for ED

The second proposition to be tested is that prices are the same fraction of market value for parcels settling in and out of court, against the alternative that high-valued parcels receive a higher price (gross of court costs) relative to market value in court awards than in voluntary settlements, because court costs or expenditures fall, relative to parcel value, more for buyers than for sellers[20]

The Southeast Englewood and Near West Side samples, for which there are data on method of settlement, were pooled and divided into subsets according to whether the price was awarded in court or agreed on in

---

[20] It may be shown that the predictions of the alternative hypothesis of the model of price determination,

$$\frac{d(P_c/P_m)}{dP_m} > 0 \quad \text{and} \quad \frac{d(P_v/P_m)}{dP_m} - \frac{d(P_c/P_m)}{dP_m} < 0,$$

imply

$$\frac{d \ln P_c}{d \ln P_m} > 1 \quad \text{and} \quad \frac{d \ln P_v}{d \ln P_m} < \frac{d \ln P_c}{d \ln P_m},$$

which is the form in which the theory was tested

TABLE 4

Court and Voluntary Subsets

Sample	Regressions of *Ped*			Regressions of $\hat{P}m$		
	C	*Ped*	R^2/SEE	*C*	$\hat{P}ed$	R^2/SEE
Total *n* = 409(1)	−8 7925 (−8 5177)	1 9443 (17 665)	4340 (0.60505)	7 2689 (60 7518)	2232 (17 6656)	4340 (0 205006)
Court *n* = 139(2)	−15 6772 (−9 20981)	2 64492 (16 9542)	.5250 (0 56203)	7 46553 (48 6196)	205416 (12 3066)	.5250 (0.19292)
Voluntary *n* = 270(3)	−4.26741 (−3 5263)	1 47504 (11 4465)	4013 (0 56203)	6 87258 (36 4998)	262747 (13 4038)	4013 (0.207576)
Court low *n* = 80(4)	−17 5926 (−4 17866)	2 86844 (6 24299)	.3330 (0 76188)	8.15226 (50 164)	116157 (6.2429)	.3330 (0 153316)
Court high *n* = 59(5)	−19 0518 (−4 80677)	3.00196 (7 27886)	4817 (0 508257)	8 0373 (37 1296)	160492 (7 278)	4816 (0 117512)
Voluntary low *n* = 146(6)	−2 31007 (−0 961430)	1 26691 (4 85109)	1404 (0 480118)	8 16297 (38 1640)	11087 (4 85109)	.1402 (0 142034)
Voluntary high *n* = 124(7)	−9 39095 (−4 4468)	2 0058 (9 0979)	3715 (0 51278)	7 75796 (38 7253)	185238 (9 0979)	.3714 (0 155829)

voluntary negotiations The results are reported in table 4 The regressions were run in both directions to estimate bounds on the true coefficient, on the assumption of additive measurement errors in both variables when the equation is estimated in logarithmic form

Comparing equations (2) and (3) in table 4 in the regressions of *Ped* on $\hat{P}_m$, we find that the intercept is lower and the coefficient on $\hat{P}_m$ higher in the court set than in the voluntary set The difference is statistically significant by a Chow test This evidence tends to refute the null hypothesis and support the alternative of greater regressivity in court A similar conclusion is implied by the regressions of $\hat{P}_m$ on *Ped* However, if these two sets of estimates are treated as bounds on the true coefficients, the ranges overlap

$$2\ 645 < \beta_c < 4\ 87, \qquad 1\ 475 < \beta_v < 3\ 81,$$

where β_c is the coefficient on $\hat{P}_m$ in the court set and β_v is the coefficient in the voluntary set The hypothesis that the true values are equal can therefore not be firmly rejected

Table 4 also reports estimates of the same equations with the sample subdivided into low- and high-valued parcels, the division being made at $12,000, which is close to the mean for both court and voluntary sets The purpose is to test the possibility that the true relation is nonlinear and that differences in the estimated coefficients for the court and voluntary sets are attributable to different frequency distributions of parcel values for the two sets The evidence does not support this conclusion

Two indirect tests of the reasonableness of the model of price determination may be performed with these estimates First, cost minimization for the buyer requires that $P_v \leq E(P_c)^b + C^b$ Both $E(P_c)^b$ and C^b are unknown However, an estimate of the lower bound on $E(P_c)^b$ is given by P_c, observed court award for a property of equal market value It is a lower bound because the voluntary set will not be a random drawing from the distribution of parcels but will tend to "select" those parcels on which the buyer overestimated the court award and thus offered a high price relative to the seller's minimum ask (assuming errors in buyer and seller expectations are not positively correlated) Then cost minimization requires that

$$P_v - P_c \leq E(P_c)^b - P_c + C^b,$$

that is, the difference between the regression estimates of settlement price and court award for a property of a given market value is a lower bound on the sum of the buyer's error in predicting court award and his court costs At the sample mean of $12,000, $P_v - P_c = \$3{,}000$, which seems not unreasonable

The second test utilizes the assumption that prices reached in voluntary

TABLE 5

SUMMARY STATISTICS OF VARIABLES FOR COURT AND VOLUNTARY SETS

	Court			Voluntary		
Variable	Mean	Antilog Mean	Variance	Mean	Antilog Mean	Variance
Ped*	9 15617	$9,475	1 00027	9 57907	14,450	0 404433
TAVL*	8 16021	3,495	0 988743	8 34385	4,250	0 590253
$\hat{P}m$*	9 35127	11,550	0 080279	9 38721	11,950	0 070708
DATE	9 64748		18 3893	7 9629		26.5488
DC	0 10791		0 09696	0 07407		0 068842
TEN	1 07914		1 986	1 8444		3 90136
VAL	75 1439		883 312	71 8333		634 080

* Arithmetic mean of logs = geometric mean of absolute numbers

	Arithmetic Mean	Variance	Arithmetic Mean	Variance
Ped	15,199 6	0 335801 E09	17,881 7	0 192506 E09
$\hat{P}m$	11,980 6	0 118874 E08	12 385 8	0 134233 E08

settlements are the average of minimum offer and maximum ask, which reduces to

$$P_v = P_c + \tfrac{1}{2}(C^b - C^s)$$

Three thousand dollars, the difference between settlement price and court award for a mean-valued property, seems excessive as an estimate of half the difference in court costs However, again the model may be salvaged by selectivity bias considerations

The voluntary set will comprise two groups of parcels, A and B, on which the buyer overestimated and the seller underestimated, respectively, P_{c_v}, the hypothetical court award on a parcel which settled out of court Assuming uncorrelated errors, for set A

$$P_v = P_{c_v} + \tfrac{1}{2}(C^b - C^s) + \tfrac{1}{2}[E(P_{c_v})^b - P_{c_v}] > \hat{P}_v,$$

and for set B

$$P_v = P_{c_v} + \tfrac{1}{2}(C^b - C^s) - \tfrac{1}{2}[P_{c_v} - E(P_{c_v})^s] < \hat{P}_v,$$

where

$$\hat{P}_v = P_{c_v} + \tfrac{1}{2}(C^b - C^s)$$

Thus, observed P_v is an unbiased proxy for $\hat{P}_v$ only if A and B are equal-sized samples from the same population and if average buyer and seller errors are equal in absolute value. If the buyer makes larger or more frequent errors, set A will dominate and observed $P_v > \hat{P}_v$, so that $P_v - P_c > \frac{1}{2}(C^b - C^s)$

The question arises whether observed P_c is an unbiased proxy for P_{c_v} By a similar argument, the court set will comprise two groups of parcels, with negative errors in buyer expectations of P_{c_v}, and therefore low offers relative to $\hat{P}_v$, and positive errors in seller expectations, and therefore relatively high asks However, even if negative buyer errors dominate in frequency or mean absolute magnitude, actual court awards are still unbiased for P_{c_v} if the courts are unaffected by the asks and offers of pretrial negotiations This assumption is plausible if it is cheap for sellers to match low offers by symmetrically high asks in court in excess of asks in pretrial negotiations This is implicit in equation (5), which makes court award only a fraction of market value and expenditure on court inputs.

Conclusions

This study of urban renewal in Chicago suggests that ED does not ensure that fair market value is paid in an assembly The mean ratio of price to market value is 1 27 for the court sample and 1 45 for the voluntary sample, with a weighted average of 1 388 This does not necessarily imply a suboptimal amount of assembly, since the opportunity cost of land for an assembly is expected to exceed mean market value However, it seems unlikely that the pattern of reservation prices corresponds to the pattern of ED prices Under ED, high-valued parcels systematically receive more than market value and low-valued parcels receive less than market value This is consistent with a simple model of how prices would be determined assuming optimizing behavior within the constraints imposed by ED and a structure of court costs which induces higher buyer expenditure relative to the seller's on low-valued properties, but the opposite relation on high-valued properties Obviously, the structure of court costs may vary in different contexts, so the conclusion with respect to the pattern of ED prices is only generalizable in similar contexts

The full-cost calculus of the relative efficiency of ED and the free market in handling assemblies cannot be made without data on comparable market assemblies and on transactions costs, including labor inputs and forgone income on land due to delay in transferring it to a higher-valued use Both components of transactions costs are likely to be higher under ED Thus, both theoretical considerations and the evidence available leave unproved the case for the superior efficiency of ED

Appendix

TABLE A1

GLOSSARY

Variable	Definition	Source
PRL	Price of property in sample of random market sales, natural logarithm	Transfer tax stamps reported in *Real Estate News*
Ped	Price of property in ED sample, natural logarithm	Unpublished HUD records
TAVL	Assessed value of land plus assessed value of improvements, natural logarithm	Cook County property tax roles: 1969–70, market sample, 1965, Lincoln Park ED sample Unpublished HUD records Southeast Englewood and Near West Side samples (date of assessment unknown)
VALL	Value per foot of frontage, natural logarithm	*Olcott's Land Values* (year prior to the year of the first observation for the market and for each ED sample)
Z	Zoning category	*Olcott's Land Values* (year as for *VALL*)
TEN	Number of tenants with listed phones	Donnelley's *Street Address Directory* (year as for *VALL*)
DC	Commercial use dummy, 0 = no commercial use, 1 = commercial use	Donnelley's *Street Address Directory*
DCT	*DC* × *TAVL*	
DATE	Number of months from month of first observation in each sample	As for *PRL* and *Ped*
DZ	Zoning dummy, 0 = more restrictive zone, 1 = less restrictive zone	As for *Z*
DZT	*DZ* × *TAVL*	
$\hat{P}_m$	Predicted market value of properties in ED sample; natural logarithm	

References

Arrow, Kenneth. "Political and Economic Evaluation of Social Effects and Externalities." In *The Analysis of Public Output*, edited by Julius Margolis. New York Nat Bur Econ. Res., 1970

Coase, R. H. "The Problem of Social Cost." *J Law and Econ.* 3 (October 1960): 1–44.

Colean, Miles L., chairman. *Report of the President's Task Force on Urban Renewal.* Washington. Government Printing Office, 1970.

George C. Olcott and Company. *Olcott's Land Values Blue Book of Chicago.* Chicago: Olcott, 1961, 1965, 1967.

Lewis, John. *A Treatise on the Law of Eminent Domain in the United States.* 3d ed. Chicago: Callaghan, 1909.

McCloskey, Donald N. "The Enclosure of Open Fields: Preface to a Study of Its Impact on the Agricultural Efficiency of Eighteenth-Century England." *J. Econ. Hist.* 32 (March 1972): 15–35.

Munch, Patricia. "An Economic Analysis of Eminent Domain." Ph.D. dissertation, Univ Chicago, 1973

Nichols, Philip. *The Law of Eminent Domain*, edited by Julius L. Sackman and Russell D Van Brunt 3d ed. rev Albany, N Y Bender, 1970.

Reuben H. Donnelley Company *Street Address Directory* Chicago Donnelley, 1961, 1965, 1967

Scheiber, Harry N. "The Road to Munn Eminent Domain and the Concept of Public Purpose in the State Courts" In *Perspectives in American History*. Vol 5, *Law in American History*, edited by Bernard Bailyn and Donald H Fleming. Cambridge, Mass.. Charles Warren Center Studies American Hist., Harvard Univ., 1971.

U S. Congress, House of Representatives, Committee on Public Works, Select Subcommittee on Real Property Acquisition *Study of Compensation and Assistance for Persons Affected by Real Property Acquisition in Federal and Federally Assisted Programs* Washington Government Printing Office, 1965

Economically Rational Expectations: Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy?

Edgar L. Feige

University of Wisconsin

Douglas K Pearce

University of Houston

The concept of "economically rational" expectation formation is developed for a regime in which the acquisition and use of some information sets are nonnegligible The concept provides a middle ground between "autoregressive" expectation formation and "rational" expectation formation A necessary condition for the use of nonnegligible cost information sets is that such sets serve as leading indicators or more formally satisfy the causality conditions of Granger (1969) Using a time series modeling identification methodology developed by Box and Jenkins (1970) and Haugh (1972), we test the causal relationship between the rate of inflation and various monetary and fiscal aggregates Surprisingly, we cannot reject the hypothesis that the rate of inflation is independent of the monetary and fiscal aggregates considered Since the proposed leading indicator series contain no incremental predictive power once the information contained in the past history of inflation is efficiently utilized, we conclude that autoregressive expectation models may indeed represent economically rational price expectation formation

Introduction

In many economic theories, anticipations of future values of economic variables play important causal roles Unfortunately, expectations held by economic agents about future economic fluctuations are generally nonobservable Thus researchers attempting to empirically evaluate these theories have been forced to find some proxy or model for expectations As Jacob Mincer (1969, p 83) has pointed out, "The issue cannot

The authors wish to acknowledge the helpful assistance of Douglas Caves

[*Journal of Political Economy*, 1976, vol 84, no. 3]

© 1976 by The University of Chicago All rights reserved

be ignored: The use of current rather than anticipated values is equivalent to a hypothesis that expectations are largely based on current magnitudes and do not differ from them in any systematic way."

In the case of inflationary expectations, the proxy employed has usually been a distributed lag model of past inflation rates. Dissatisfaction with this approach has led to recent work (Modigliani and Shiller 1973; Sargent 1973, Sargent and Wallace 1973) which has utilized the rational expectations concept of Muth (1961). Rational inflationary expectations are defined by Muth to be equivalent to the forecasts of inflation from the economic model deemed appropriate by the agent forming the expectations

This paper develops a different approach to expectation formation, focusing on the efficient use of information in forecasting inflation In Section I, we discuss the types of information economic agents might utilize in forming price expectations and introduce the notion of "economically rational" expectations This latter concept simply emphasizes that an economic agent should consider the trade-off between the benefits and costs of added information when forecasting, and thus anticipating, future inflation rates The "rational expectations" framework focuses attention on the structural model generating the inflation series, but it implicitly assumes that the cost of acquiring and utilizing such information is zero Where information costs are nontrivial, the economically rational individual may opt for a less expensive forecasting framework even at the expense of a larger standard error of forecast One relatively inexpensive method of forecasting inflation is to consider only the information contained in past inflation rates Section II explores the problem of identifying and estimating an efficient forecasting model for inflation when no information other than past inflation is considered Using the time series methodology of Box and Jenkins (1970), we estimate such forecasting models for U S data and show that they can be viewed as generalizations of the adaptive expectations approach popularized by Cagan (1956) In Section III, we consider the conditions under which it would be beneficial to employ a leading indicator series to increase forecasting accuracy It will be shown that for a series to qualify as a leading indicator in this framework, the series must "cause" inflation in the sense of Granger (1969) We then examine measures of monetary and fiscal policy to see if they can be considered as leading indicators of inflation We find that the incremental information contained in these measures is not useful for predicting inflation given that efficient use has already been made of the information contained in past inflation rates The final section presents our conclusions

I. Information and Economically Rational Expectations

Following Hicks (1946) we categorize three types of information upon which price expectations could be formed· (*i*) noneconomic events such as the weather, politics, wars, and psychological variables; (*ii*) structural economic events such as changes in monetary and/or fiscal policy; and (*iii*) the "inflation experience"—the actual past history of price changes.

The manner in which elements from the first information set affect inflationary anticipations is not well known.[1] Such events are typically regarded by economists as stochastic shocks which do not produce a predictable systematic effect

Elements from the second information set would be relevant to an agent forming rational expectations in the sense of Muth since these expectations "are essentially the same as the predictions of the relevant economic theory" (Muth 1961, p 316) If one's "relevant theory" is that changes in the quantity of money affect the rate of inflation with some distributed lag, it would seem plausible that one's expectations of future inflation rates should be influenced by the history of changes in the money supply Muth does not, however, explicitly consider the costs of finding and making operational the relevant theory

The third information set contains the past history of actual inflation rates Practically all the empirical work examining the effects of inflationary anticipations has employed this information set by using a distributed lag model The general model is

$$p_t^e(l) = \sum_{i=0}^{n} w_{il} p_{t-i}, \tag{1}$$

where $p_t^e(l)$ = expected inflation rate for time $t + l$ formed at time t and p_{t-i} = actual inflation rate at time $t - i$ The familiar adaptive expectations model of Koyck (1954) and Cagan (1956) treats $p_t^e(l)$ as equal for all l That is, the agent forecasts a single inflation rate for all future time periods As was shown by Muth (1961) this type of forecast is optimal if the process generating the inflation rate is given by

$$p_t = p_{t-1} + a_t - \theta_1 a_{t-1} \tag{2}$$

where a_t is a random variable with

$$\begin{aligned} E(a_t) &= 0 \\ E(a_t a_s) &= \sigma_a^2, \qquad t = s \\ &= 0, \qquad t \neq s \end{aligned}$$

If we represent the forecast rate as a weighted sum of past rates as in equation (1), this model constrains the weights to fall off exponentially [2]

---

[1] Surveys of consumer attitudes following events included in information set (i) have been conducted by the Survey Research Center of the University of Michigan See Katona (1972) for a summary of the results The effects of such events on stock prices have been examined by Niederhoffer (1971) and Hester (1972)

[2] A model which is slightly more general but still consistent with Koyck and Cagan allows for a deterministic trend by including a constant in equations (1) and (2) In this case, the forecast is not the same for all l but is a linear function of l

Papers by Walters (1971) and Sargent and Wallace (1973) have also worked with models in which expectations are based only on the inflation experience. They conclude that such expectations are rational (Walters prefers the term "consistent") if the underlying structural model of the inflation process can be collapsed into a reduced form equation with past inflation rates as the predetermined variables.

This class of models is, however, a very small subset of the general class of multivariate time series processes. As is shown in Zellner and Palm (1974) a multivariate time series process will almost always yield a set of univariate equations in which each variable can be expressed as a function of its own past values. Expectations formed on the basis of these univariate equations will in general have a larger mean squared error than expectations which are rational in the sense that they employ all the available information in the model (Nelson 1975) Thus, while the potential benefits of utilizing all available information are apparent, the absence of an explicit consideration of the information costs which would be incurred in forming rational expectations is a serious drawback We, therefore, suggest the notion of economically rational expectation formation. An economic agent is economically rational only if he considers both the costs of misestimating future inflation and the costs of making his forecast of future inflation

Assume the costs of misestimating inflation can be adequately represented by a quadratic loss function[3] such as

$$C = k\ (p_t - p^e_{t-1}[1])^2 \tag{3}$$

Further assume that the squared forecast error is a function of the amount and type of information employed such that

$$(p_t - p^e_{t-1}[1])^2 = f(I_1, \ldots, I_m) \qquad f'_{I_i} > 0, \quad f''_{I_i} < 0, \tag{4}$$

where I_i = ith type of information available at price c_i

The total cost of misestimating inflation is thus.

$$\begin{aligned} TC &= k \cdot (p_t - p^e_{t-1}[1])^2 + \sum_{i=1}^{m} c_i I_i, \\ &= k \cdot f(I_1, \ldots, I_m) + \sum_{i=1}^{m} c_i I_i \end{aligned} \tag{5}$$

---

[3] In many cases the loss function may be asymmetrical since for certain economic agents underestimating inflation is much more costly than overestimating inflation.

The cost minimizing action is of course to buy units of each type of information until the reduction in the cost of the forecast error per dollar spent on each type of information is equated, such as:

$$c_i = -k \cdot f_i' \qquad i = 1, m. \tag{6}$$

The specific costs of different types of information are difficult to estimate and may vary across individuals and across time. We make the strong assumption that the information cost on past inflation rates is relatively small and thus this information should be utilized intensively. In the next section we identify and estimate expectation models which efficiently use all the predictive information contained in the past history of inflation In Section III we then look at the additional predictive benefits of also taking into account changes in monetary and fiscal policy.

II Efficient Autoregressive Expectation Models

We seek an expectation model, based on past inflation rates, which utilizes all the predictive information contained in the inflation rate series The time series methodology of Box and Jenkins (1970) is ideally suited to finding such a model This approach makes the basic assumption that the time series under observation is a realization from a stochastic process The problem then becomes one of identifying and estimating the underlying process

The stochastic process generating the inflation rate series can be represented as a weighted sum of random shocks [4]

$$\begin{aligned} p_t &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \qquad (7) \\ &= \psi(B) a_t, \end{aligned}$$

or under very general assumptions as a distributed lag of past inflation rates plus a random error (Whittle 1963)

$$p_t = \pi_1 p_{t-1} + \pi_2 p_{t-2} + \quad + a_t \tag{8}$$

or simply, $\pi(B)p_t = a_t$, where p_t = inflation rate at time t, $\psi(B)$ and $\pi(B)$ = polynomials in the lag operator B (i e , $B^i x_t = x_{t-i}$), a_t = random error with $E(a_t) = 0$, $E(a_t a_{t-s}) = \sigma_a^2 I$

In order to avoid the problems involved in estimating models of the form of (7) and (8), Box and Jenkins employ a general class of time series models known as autoregressive integrated moving average (ARIMA) models. The general ARIMA model has the form

$$\phi_p(B)(1 - B)^d p_t = \theta_q(B) a_t$$

---

[4] That such a linear decomposition exists for any stationary stochastic process was proved by Wold (1954).

where $\phi_p(B)$ and $\theta_q(B)$ are polynomials in B with roots outside the unit circle and $(1 - B)^d$ is the dth difference of the series.[5] The time series, appropriately differenced to achieve stationarity, is thus expressed as a mixed autoregressive (of order p) moving average (of order q) process. The order of the model and its pattern of seasonality are identified by means of the autocorrelation function. The identified model is then efficiently estimated[6] and subjected to two diagnostic checks on the residuals First, the individual autocorrelations are compared to their estimated standard errors Second, an overall test of the adequacy of the model is made by calculating the statistic Q.

$$Q = (N - d) \sum_{k=1}^{T} r_k^2(\hat{e}_t)$$

where N = number of data observations, d = number of differences required for stationarity, and $r_k^2(\hat{e}_t)$ = square of the estimated autocorrelation coefficient of $\hat{e}_t$ and $\hat{e}_{t-k}$ This statistic is approximately distributed as $\chi^2(T - p - q)$ Model inadequacy would be indicated if the calculated value of Q exceeded the appropriate χ^2 value

The above procedures were applied to monthly and quarterly inflation series as measured by the rates of change in the Consumer Price Index (CPI) and the Wholesale Price Index (WPI) for U S unadjusted data from mid-1953 to mid-1971 [7] Table 1 presents the estimated ARIMA models for these series

The ARIMA model identified and estimated for the monthly rate of change in the CPI is somewhat troublesome since a seasonal pattern with a periodicity of 4 months is indicated It is difficult to think of an economic reason why this periodicity should occur [8] Since this model is similar to that estimated for the rate of change in the WPI, except for the seasonality terms, we estimated a corresponding model for the CPI with the following result

$$(1 - B)p_t = (1 - 845B)a_t \qquad Q = 54\ 85$$
$$\pm\ 072 \qquad (39\ \text{df})$$

This model is less satisfactory in terms of the diagnostic tests on the estimated residuals and so the model listed in table 1 was retained

Note that only in the case of the WPI monthly data does the model

---

[5] Differencing is required if the original series is nonstationary.

[6] The estimation procedure yields approximate maximum likelihood estimates if the errors are assumed to be normally distributed

[7] The data source for the price indices is *1971 Business Statistics* (see U S Dept of Commerce 1972).

[8] Breaking the time period into two equal parts and repeating the identification and estimation procedure for each subset yielded almost identical models

TABLE 1

ESTIMATED ARIMA MODELS FOR INFLATION SERIES
(95% CONFIDENCE INTERVAL)

	Monthly Data 1953(7)–1971(6)	
CPI	$(1 - B^4)(1 - B)p_t = (1 - .567B)(1 - .766B^4)a_t$ $\pm .118 \quad \pm .090$	$Q = 44.45$ (38 df)
WPI	$(1 - B)p_t = (1 - .765B)a_t$ $\pm .086$	$Q = 41.16$ (39 df)
	Quarterly Data 1953(III)–1971(II)	
CPI	$(1 - B^4)(1 - .649B)p_t = .029 + (1 - .822B^4)a_t$ $\pm .179 \quad \pm .020 \quad \pm .166$	$Q = 28.70$ (37 df)
WPI	$(1 - B^4)(1 - .300B)p_t = .039 + (1 - .693B^4)a_t$ $\pm .231 \quad \pm .028 \quad \pm .189$	$Q = 28.77$ (37 df)

correspond to the specific class of models studied by Koyck and Cagan.[9] For the WPI monthly model, forecasts made at time t will be the same for all lead times and this forecast is arrived at by taking an exponentially weighted average of past inflation rates as in equation (1).

The implied weights placed on past inflation rates, the π weights of equation (8), can be derived once the ARIMA models have been estimated. The ARIMA representation is

$$\phi(B)(1 - B)^d p_t = \theta(B)a_t$$

or

$$\frac{\phi(B)(1 - B)^d}{\theta(B)} p_t = a_t \tag{9}$$

This implies that, from equation (8),

$$\begin{aligned} \frac{\phi(B)(1 - B)^d}{\theta(B)} &= \pi(B) \\ &= 1 - \pi_1 B - \pi_2 B^2 - \end{aligned} \tag{10}$$

Since the parameters of $\phi(B)$ and $\theta(B)$ have been estimated and the appropriate degree of differencing, d, has been identified, estimates of the π weights can be derived by equating the coefficients on different powers of B in equation (10).

The π weights implied by the estimated ARIMA models are plotted in figure 1. With the exception of the WPI model, the exponential decline of weights implied by Cagan's adaptive expectations model does not occur.[10]

---

[9] The weights are $w_i = (1 - \theta)\theta^i$ where θ is the moving average parameter (.765 in the WPI model).

[10] It should be noted that the actual weights are determined by a forecasting criterion in this paper. In Cagan's application, the weights are determined by a least-squares estimate of a demand for money model.

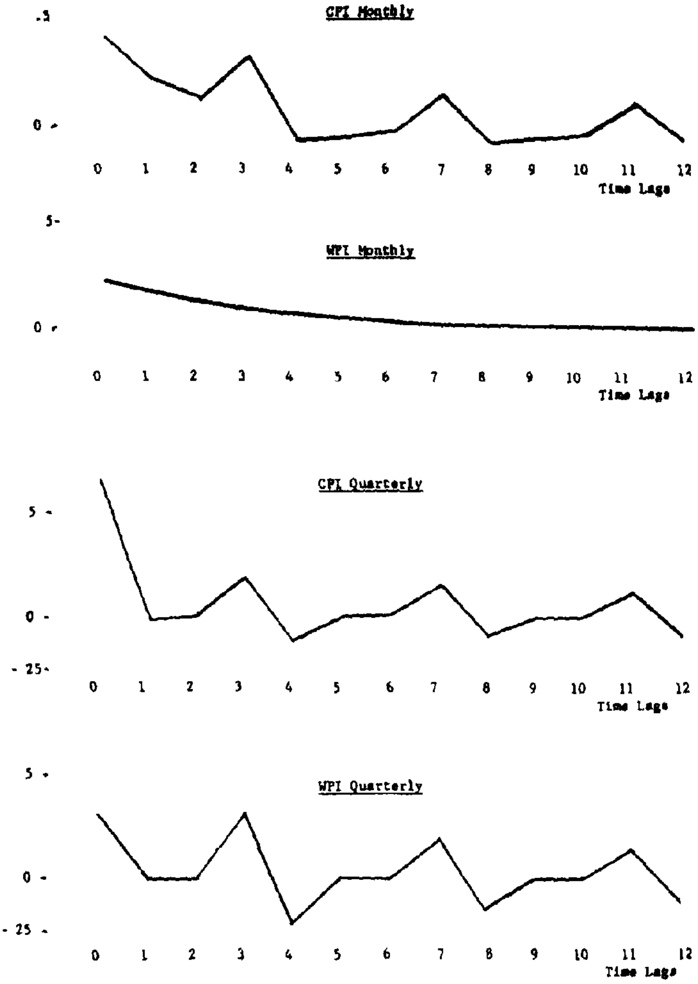


FIG 1 —Weights placed on past inflation rates implied by ARIMA models

The one-step ahead forecast can be calculated from the autoregressive models as follows

$$p_t^e(1) = \pi_1 p_t + \pi_2 p_{t-1} + \pi_3 p_{t-2} + \cdot$$

where $p_t^e(1)$ = the forecasted inflation rate for time $t + 1$ made at time t. The forecast is simply the conditional expectation

$$E(p_{t+1} \mid p_t, p_{t-1}, p_{t-2}, \quad .) = p_t^e(1).$$

Since $p_{t+1} = a_{t+1} + \pi_1 p_t + \pi_2 p_{t-1} + \cdots$ and thus $E(p_{t+1}) = \pi_1 p_t + \pi_2 p_{t-1} + \cdots$, the one-step ahead forecast error $e_t(1)$ will be

$$\begin{aligned} e_t(1) &= p_{t+1} - p_t^e(1) \\ &= a_{t+1} \end{aligned} \tag{11}$$

Thus the random shocks driving the stochastic process can be viewed as the one-step ahead forecast errors

As shown by Box and Jenkins (1970), the forecasting models can also be viewed as adaptive or error-learning expectation models Recall that the stochastic process generating the inflation series can be represented as a weighted sum of current and past random shocks

$$\begin{aligned} p_t &= a_t + \sum_{i=1}^{\infty} \psi_i a_{t-i} \\ &= \psi(B) a_t, \end{aligned} \tag{12}$$

thus

$$p_{t+l+1} = a_{t+l+1} + \psi_1 a_{t+l} + \cdots$$

The l-step ahead forecast made at time $t + 1$ is the conditional expectation

$$p_{t+1}^e(l) = E_{t+1}[p_{t+l+1} \mid a_t, a_{t-1}, \ldots] = \psi_l a_{t+1} + \psi_{l+1} a_t + \cdots$$

Similarly,

$$p_t^e(l + 1) = \psi_{l+1} a_t + \psi_{l+2} a_{t-1} + \cdots$$

thus

$$p_{t+1}^e(l) = p_t^e(l + 1) + \psi_l a_{t+1} \tag{13}$$

Equation (13) implies that the expected inflation rate for some period in the future, in this case the period is $t + l + 1$, is revised when new information appears At time $t + 1$, the economic agent experiences an error, equal to a_{t+1}, in his forecast of inflation for that period (see eq [11]) He then revises all his previous expectations of future inflation rates by $\psi_l a_{t+1}$ Because the weights ψ_l depend on l, the proportion of the currently experienced error he adds to his old expectations varies with the horizon of the expectations In the special case of the adaptive expectations model of Cagan, the ψ_l weights are identical over all horizons.[11] The patterns of the ψ weights for the estimated ARIMA models are given in figure 2

---

[11] In terms of an ARIMA model, the Cagan model is represented as $(1 - B)p_t = (1 - \theta B)a_t$ Thus $\psi(B) = (1 - \theta B)/(1 - B)$ This yields $\psi_i = 1 - \theta$, for all i

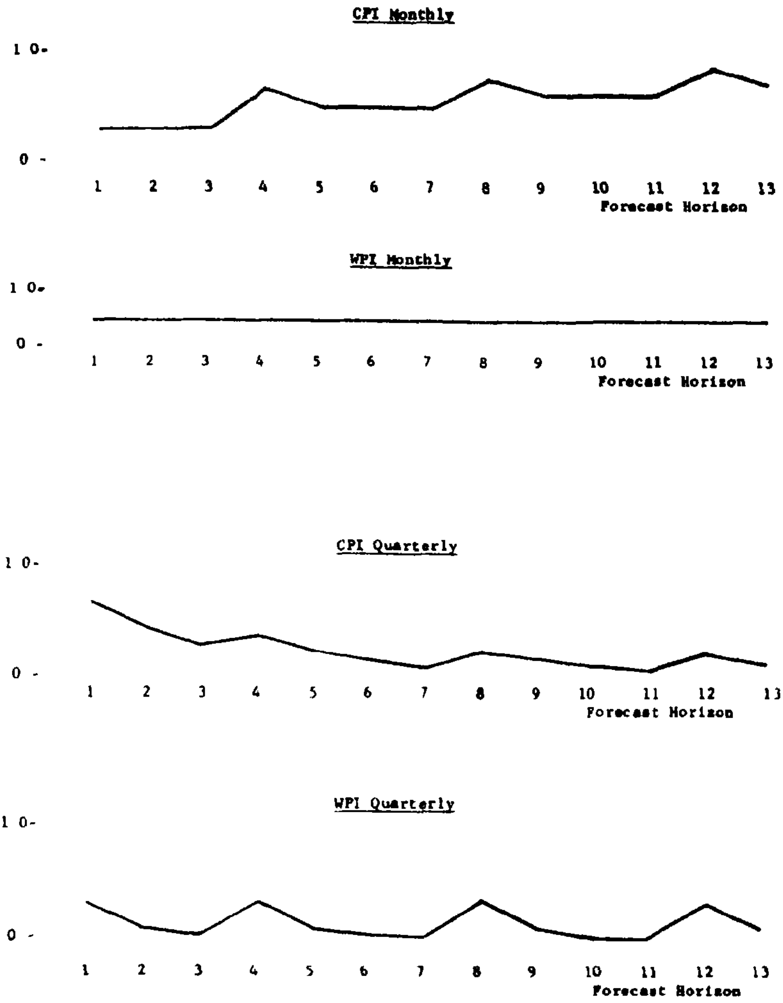


FIG 2 —Adaptive expectations weights for different horizons

Thus the forecasts from the ARIMA models listed in table 1 can be viewed as optimally adaptive expectations given that the economic agent utilizes only the information contained in past inflation rates when forming his expectations

III. The Use of Leading Indicators of Inflation

Having found models which efficiently utilize the information contained in the inflation experience, we now ask the question, How can one evaluate the predictive benefits from enlarging the information set to include other economic time series? It is shown below that this question is empirically identical to the issue of assessing the causal relation, as defined by Granger (1969) between two times series. One procedure for studying causal relationships has been developed by Haugh (1972) building on the work of Box and Jenkins (1970)

Since an ARIMA model for inflation efficiently utilizes all the information contained in the inflation series, a necessary condition for a series to serve as a leading indicator is that it provide additional information relevant to future inflation, that is, that it reduce the expected mean square forecast error below that of the models based solely on past inflation rates For an economically rational agent, this criterion is not sufficient, since he will use the leading indicator only if the predictive benefits outweigh the costs

More formally, a necessary condition for a time series, x, to be a leading indicator of the inflation series, p, is.

$$\sigma^2(p \mid \bar{p}, \bar{x}) < \sigma^2(p \mid \bar{p}) \tag{14}$$

where $\sigma^2(p \mid \bar{p}, \bar{x})$ is the minimum predictive error variance of inflation given both past inflation rates (denoted $\bar{p}$) and the past values of the x series (denoted $\bar{x}$), and $\sigma^2(p \mid \bar{p})$ is the minimum predictive error variance of inflation given past inflation rates

Condition (14), however, corresponds to the condition set forth by Granger (1969) under which one can say x "causes" p Granger's empirically testable definition of causality states that x causes p if past values of x can be used to improve the conditional predictions of future p

How does one test for this causal relation?[12] The residuals from the ARIMA models of inflation can be viewed as that part of inflation which cannot be predicted from its own past history Thus in order for series x to enable one to improve the forecasts of inflation from the ARIMA models, it must be correlated with the residual series from the ARIMA models Haugh has proposed the following procedure for analyzing this condition First, ARIMA models must be identified and estimated for the p series and the prospective leading indicator series x[13] Second, the residuals from these models are calculated—the residual series being referred to as the *prewhitened series* Third, the prewhitened inflation series is cross-correlated with the prewhitened leading indicator series An examination of the sample cross-correlation function should indicate the existence and direction of causality in the sense of Granger (1969) The reason for not simply cross-correlating the prewhitened inflation series with the leading indicator series is that the resulting cross-correlation function is often misleading due to the serial correlation present in the leading indicator series (Bartlett 1935)

---

[12] For another approach to empirically applying Granger's causality criteria, see Sims (1972)

[13] It is assumed that the x series is also a realization of a stochastic process.

Suppose we have the population univariate disturbances u_{x_t} and u_{p_t} for the x and p series Let $\rho_{u_x u_p}(k)$ be the correlation between u_{x_t} and $u_{p_{t+k}}$ It is shown by Haugh (1972) that if $\rho_{u_x u_p}(k) \neq 0$ for some $k > 0$, then condition (14) is satisfied and we can say "x causes p" Similarly, if $\rho_{u_x u_p}(k) \neq 0$ for some $k < 0$, p causes x If there are nonzero $\rho_{u_x u_p}(k)$ for both positive and negative k, there is a feedback relationship, and if $\rho_{u_x u_p}(k) = 0$ for all k, x and p are independent The case of contemporaneous cross-correlation is ignored for, in this case, the direction of causality is indeterminate Thus, the causality relationship between two variables can be assessed by examining the cross-correlation between the residuals from the estimated univariate models

Let $r_{\hat{u}_x \hat{u}_p}(k)$ be the sample correlation between the residual $\hat{u}_{x_t}$ from the prewhitened x and $\hat{u}_{p_{t+k}}$ from the prewhitened p series Once the sample cross-correlation function is obtained, a sequential process is employed to determine the causal relationship First, the hypothesis of independence is considered and subjected to two statistical tests On the null hypothesis of independence, the sample cross-correlations, $r_{\hat{u}_x \hat{u}_p}(k)$, are each distributed normally with mean zero and standard deviation $1/\sqrt{N}$, N being the length of the sample time series One way to proceed is to compare each $r_{\hat{u}_x \hat{u}_p}(k)$ to its standard error $1/\sqrt{N}$, noting the number and position of those $r_{\hat{u}_x \hat{u}_p}(k)$ exceeding, say, two standard deviations The appearance of a number of cross-correlations significantly different from zero by this measure would indicate a lack of independence The location of nonzero $r_{\hat{u}_x \hat{u}_p}(k)$ is indicative of the direction of causality in the relationship

A second test for independence is to compute the statistic

$$S = N \sum_{k=-m}^{m} r^2_{\hat{u}_x \hat{u}_p}(k),$$

which on the null hypothesis is the sum of $2M + 1$ squared standard normal variables and is distributed asymptotically as $\chi^2(2M + 1)$ If the computed S exceeds the $\chi^2(2M + 1)$ variable at the chosen significance level, the hypothesis of independence is rejected

Since independence rules out both feedback and unidirectional causality in either direction, the testing procedure is terminated if the hypothesis of independence is not rejected If independence is rejected, the sample cross-correlation function may be suggestive of the type of causality that is present, but formal hypothesis tests involve more complex procedures requiring the estimation of a multivariate model, or the use of the Sim's (1972) regression procedure As is shown below such further tests are unnecessary in the present application

We have applied this approach to see if certain measures of monetary and fiscal policy could be considered useful leading indicators of inflation. Again, we emphasize that this is only a necessary and not a sufficient condition for an economically rational agent to employ these series when forming his inflationary expectations. The measures of monetary policy chosen were the rates of change in monetary aggregates, namely: currency plus demand deposits ($M1$), currency, demand deposits, and time deposits ($M2$), and the monetary base (MB). The measure of fiscal policy selected was the high (or full) employment budget surplus (FP).[14]

The first step in the procedure is the estimation of ARIMA models for the prospective leading indicator series. Table 2 presents the estimated models for both monthly and quarterly data. The residuals from each model passed the diagnostic tests described above.

As table 2 indicates, the monthly rates of change in the monetary aggregates exhibited a seasonal pattern with a periodicity of 12 months. The quarterly rates of change also had a pronounced seasonal component with a periodicity of 4 quarters. There appeared to be no systematic seasonal pattern in the high employment budget series.

Having estimated suitable ARIMA models for both the inflation series and leading indicator series, and thus obtained the residual series or prewhitened series, the second step in the Haugh procedure can be taken. Each prewhitened inflation series (i.e., that part of inflation that could not be predicted from the past history of inflation) was cross-correlated with each prewhitened leading indicator series. These sample cross-correlation functions are illustrated in figures 3 and 4.

Two tests were employed in the examination of the sample cross-correlation functions. First, individual cross-correlations were compared to their approximate standard errors. Second, an overall test for independence between the series was performed using a slight modification of the S-statistic presented above. On the null hypothesis, S is distributed asymptotically as $\chi^2(2M + 1)$, but Haugh (1972) showed that in small samples, the $\chi^2(2M + 1)$ distribution is more closely approximated by the modified statistic

$$S^* = N^2 \sum_{k=-m}^{m} (N - k)^{-1} r^2_{\hat{a}_x \hat{a}_p}(k)$$

where N = number of observations on each prewhitened series, M = number of cross-correlations estimated for both negative and positive lags, $r^2_{\hat{a}_x \hat{a}_p}(k)$ = square of the estimated cross-correlation coefficient at lag

---

[14] The monetary series were seasonally unadjusted and available both monthly and quarterly. The high employment budget surplus was only available quarterly and was seasonally adjusted. Data sources were: $M1$, $M2$, MB – *Federal Reserve Bulletin*; FP— "Technical Notes for Estimates of the High-Employment Budget," Federal Reserve Bank of St. Louis, March 1973, mimeographed.

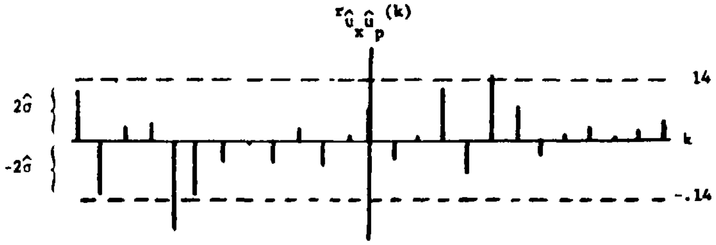
TABLE 2

ARIMA Models for Leading Indicator Series (95% Confidence Interval)

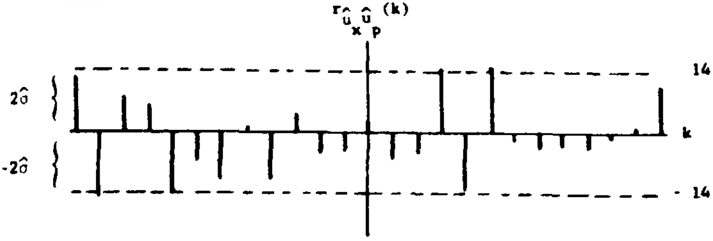
	Monthly 1953(7)–1971(6)	
*M*1	$(1 - B^{12})M1_t = (1 + \underset{\pm 142}{182}B^2)(1 + \underset{\pm 126}{276}B^3 - \underset{\pm 130}{318}B^{12})a_t$	$Q = 37\ 27$ (37 df)
*M*2	$(1 - B^{12})(1 - \underset{\pm 142}{375}B^3)M2_t = (1 + \underset{\pm 141}{328}B + \underset{\pm 143}{290}B^2)(1 - \underset{\pm 130}{489}B^{12})a_t$	$Q = 25\ 43$ (36 df)
MB	$(1 - B^{12})MB_t = \underset{\pm 020}{026} + (1 - \underset{\pm 116}{628}B^{12})a_t$	$Q = 42\ 37$ (38 df)
	Quarterly 1953(II)–1971(II)	
*M*1	$(1 - B^4)(1 - B)(1 - \underset{\pm 224}{705}B + \underset{\pm 270}{291}B^2)M1_t = (1 - \underset{\pm 197}{835}B)(1 - \underset{\pm 278}{535}B^4)a_t$	$Q = 22\ 42$ (36 df)
*M*2	$(1 - B^4)M2_t = \underset{\pm 083}{093} + (1 + \underset{\pm 236}{893}B + \underset{\pm 254}{377}B^2)(1 - \underset{\pm 186}{768}B^4)a_t$	$Q = 16\ 45$ (36 df)
MB	$(1 - B^4)\ MB_t = \underset{\pm 036}{077} + (1 - \underset{\pm 256}{310}B^7)(1 - \underset{\pm 261}{613}B^4)a_t$	$Q = 22.29$ (37 df)
FP	$(1 - B)FP_t = (1 - \underset{\pm 255}{314}B^8)a_t$	$Q = 18.27$ (39 df)

Note.—*M*1, *M*2 and *MB* are percentage rates of change

M1 and CPI:



M2 and CPI



MB and CPI

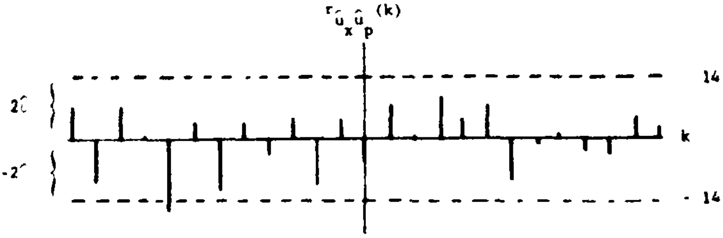
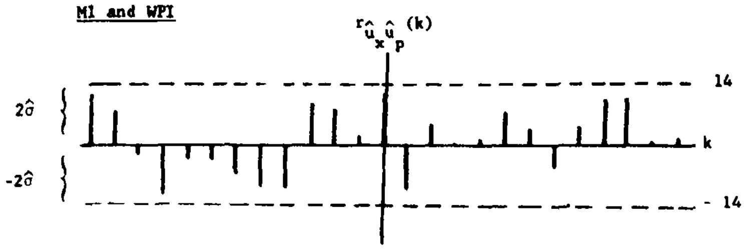
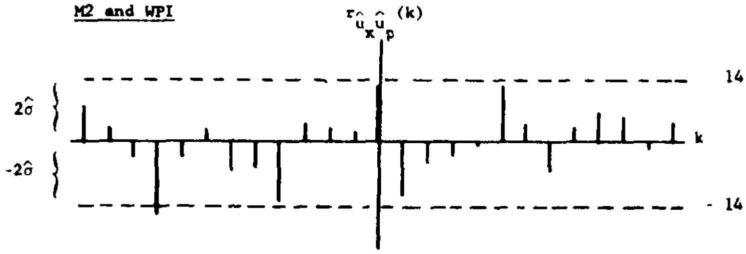


FIG 3 Cross-correlation functions for prewhitened inflation and leading indicator series, monthly data 1953(7)–1971(6)

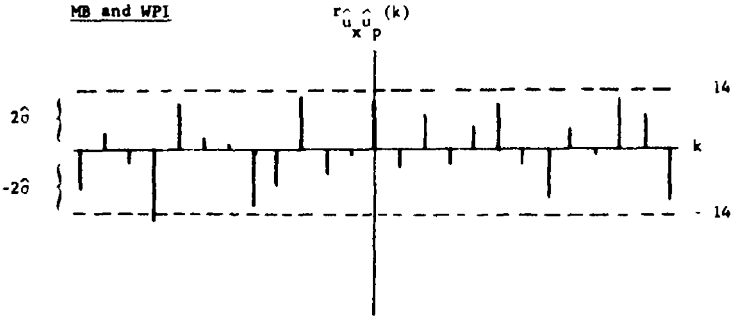
(continued overleaf)



M1 and WPI



M2 and WPI



MB and WPI

FIG 3 — *continued*

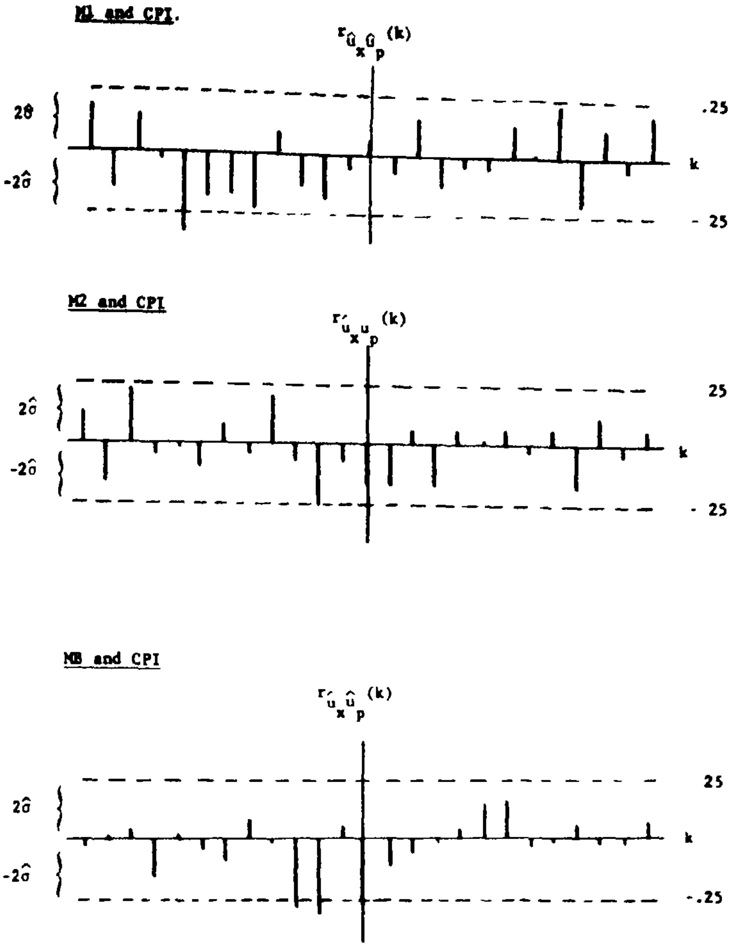
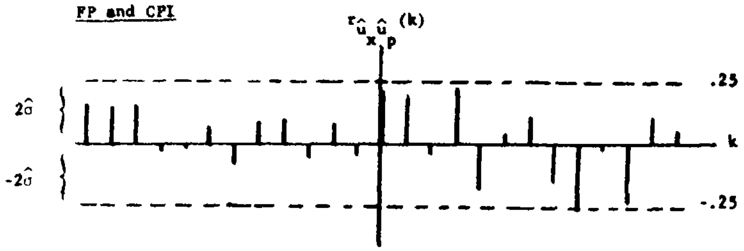
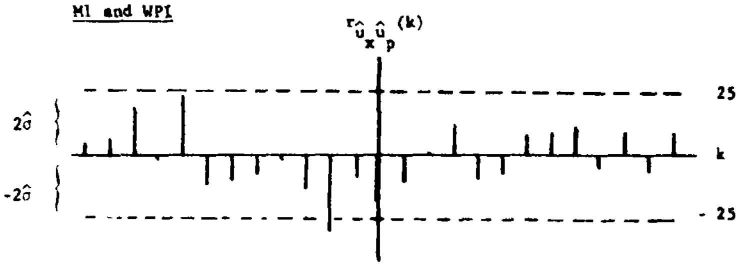


Fig 4 Cross-correlation functions for prewhitened inflation and leading indicator series, quarterly data 1953(III) 1971(I) *(continued overleaf)*

FP and CPI



M1 and WPI



M2 and WPI

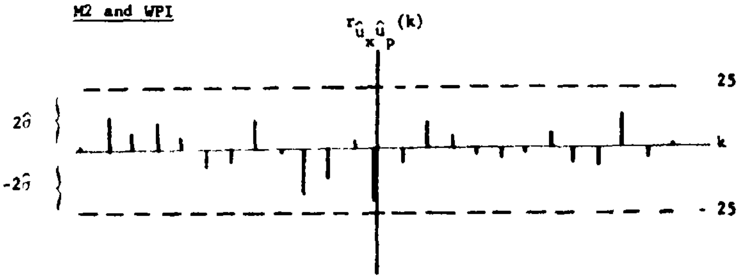


FIG 4—*continued*

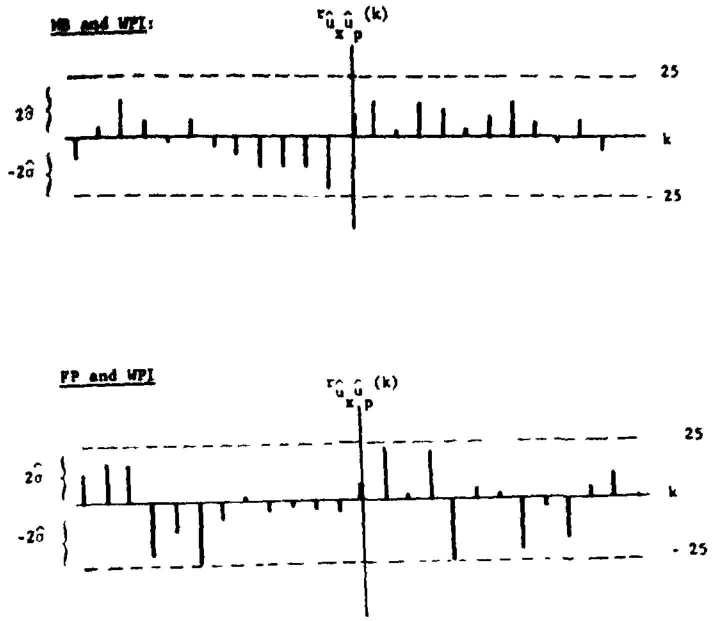


FIG 4—*continued*

k for the prewhitened leading indicator series (x) and the prewhitened inflation series (p). The hypothesis of independence between the two series would be rejected if the calculated S^*-value exceeded the tabled χ^2 value at the selected significance level

Examination of the individual cross-correlations for the estimated functions revealed very few significant spikes, and no patterns which would suggest causality from monetary or fiscal aggregates to inflation. More formal tests of independence are provided by the S^*-statistics presented in table 3 The values indicate that for no combination of leading indicators and inflation rates can the hypothesis of independence be rejected at the 5 percent level of significance[15]

To determine whether the test outcome is at all sensitive to the selection of particular values of M (where M is the maximum lag used in the construction of S^*), we computed S^* for all values of M from one to 30. The results indicated that the test was insensitive to the selection of M and thus we present only those statistics based on $M = 30$ for monthly data and $M = 12$ for quarterly data

---

[15] Similar results were obtained when the simpler ARIMA model for the CPI was used to prewhiten the CPI inflation series.

TABLE 3

S^*-STATISTICS FOR TESTING THE HYPOTHESIS OF NO RELATIONSHIP BETWEEN INFLATION MEASURES AND POSSIBLE LEADING INDICATORS

LEADING INDICATOR	INFLATION MEASURE	
	CPI	WPI
Monthly Time Series 1953(7)–1971(6)		
*M*1	56.87	48.64
*M*2	67.93	50.57
MB	69.25	55.40
Quarterly Time Series 1953(III)–1971(II)		
*M*1	34.97	23.60
*M*2	25.23	15.73
MB	20.62	14.90
FP	22.25	23.52

NOTE.—For monthly time series, S^* based on M, maximum lag, of 30, $\chi^2(61) = 79.1$ at 5% level. For quarterly time series, S^* based on $M = 12$, $\chi^2(25) = 37.7$ at 5% level. *M*1, *M*2, and *MB* are percentage rates of change.

This evidence strongly suggests that the information contained in the monetary and fiscal aggregate series would not yield more accurate predictions of inflation than an efficient autoregressive model. Therefore, for this sample period, neither monetary aggregates nor the high employment surplus can be considered as leading indicators of inflation, and an economically rational agent would not employ these series to help him forecast inflation.

Summary and Conclusions

If we are to improve our ability to predict and control inflation, it will be necessary to gain a greater understanding of the mechanisms by which inflationary expectations are generated. Applied empirical workers have relied almost exclusively on autoregressive models in efforts to specify expectation generating processes, whereas more recent theoretical work has focused attention on rational expectation formation. The autoregressive or "adaptive" models specify that expectations are generated solely on the basis of the past history of inflation. Alternatively, rational expectation models require that the public's expectations are equivalent to the mathematical expectation of inflation based upon knowledge of the model generating inflation as well as the available information pertinent to the model. Autoregressive specifications have been criticized as being behaviorally too naive, yet they have the great advantage of empirical tractability. Rational expectation models offer the theoretical appeal of greater consistency with the economist's paradigm of rational behavior in a world of negligible information costs, however, they avoid speaking to the empirical issue of selecting a particular relevant information set.

We attempt to establish a middle ground between the two approaches by introducing the concept of economically rational expectations which are formed by balancing the predictive benefits derivable from alternative information sets against the costs of acquiring and utilizing alternative information sets In order to give economically rational expectations empirical content, we begin with the strong assumption that the cost associated with the acquisition and utilization of the information set containing the past history of inflation is negligible compared to the predictive benefits to be gained from using such information intensively Under such conditions, the economic agent would initially generate expectations by forecasting with an autoregressive model which can be viewed as an adaptive or error learning process in which expectations are revised in an optimal manner Autoregressive forecasting models which efficiently utilize the systematic information contained in the past history of inflation were estimated from U S data using Box-Jenkins time series procedures

We then considered the conditions under which it would be economically rational to utilize additional information sets for forecasting purposes Assuming that costs of incremental information are non-negligible, the necessary condition for utilizing an additional time series in forecasting is simply that the time series serves as a leading indicator in the sense of incrementally improving on the forecasts of the autoregressive models This necessary condition is formally identical to Granger's (1969) causality condition, and can be empirically tested by examining the cross-correlation function of the "innovations" in the inflation series and the innovations in the proposed leading indicator series We selected several monetary aggregates and a measure of fiscal policy as the most likely candidates for containing information which would be useful in improving forecasts of inflation To our surprise, the empirical tests indicated that we could not reject the hypothesis that the rate of inflation is causally independent of the monetary aggregates and the measure of fiscal policy The results indicated that the proposed leading indicator series contained no additional predictive power over and above the information already contained in the past history of the inflation series Thus, for the information sets examined, the results suggest that forecasts from efficient autoregressive models can be considered economically rational expectations

The inability to reject the hypothesis of independence between innovations in the rate of inflation and innovations in aggregative measures of monetary and fiscal policy, while lending support for the common usage of autoregressive expectation specifications, nevertheless leaves us with a disturbing bit of evidence which appears to be in direct conflict with both popular doctrine and a substantial body of published econometric literature Our anomalous results might simply be attributable to the usual and unusual caveats associated with the fledgling methodology employed in this study;[16] however, we would regard it as inappropriate to dismiss the findings of "independence" so lightly. Two important developments persuade us that these results should be given greater credence First, there are a number of recent empirical studies corroborating the "independence phenomenon" in areas where economists would generally expect to find clear evidence of causal relationships (Cramer and Miller 1974; Feige and Pearce 1974; Pearce 1974; and Pierce 1974) In a similar vein, recent forecasting experiments with ARIMA models suggest that once Box-Jenkins methods have been applied to efficiently utilize the systematic information content of the past history of an economic variable, there is relatively little incremental information to be gained even from such complex sources as large scale econometric model forecasts (Naylor, Seaks, and Wichern 1972, Nelson 1972) The second development which shapes our perspective is the growing realization that much of the published empirical literature based on aggregative time series data may be contaminated with a proliferation of Type I errors (Box and Newbold 1972, Feige 1975, Granger and Newbold 1974, Lovell 1974) Given the increased awareness of the "spurious regression" phenomenon and the demonstration that our all too common "data grubbing" procedures produce "statistically significant" relationships between independent variables, perhaps the time has come to give greater attention to results which conflict with our received paradigm Such results will encourage both a reformulation of the theoretical underpinnings of our macroanalysis, and perhaps more significantly, a reexamination of the adequacy of our aggregative time series data as a basis for verifying and testing propositions from economic theory (Pierce 1974) The historical laboratory on which we so heavily depend for the generation of data to test our theories may simply perform nugatory experiments of promiscuous design. The challenge then is to discover a more viable affinity between the statement of economic hypotheses and the data base appropriate for testing them

References

Bartlett, M. S "Some Aspects of the Time-Correlation Problem in Regard to Tests of Significance " *J Royal Statis Soc.*, pt 3 (1935), pp 536–43.

Box, G. E. P., and Jenkins, G M *Time-Series Analysis Forecasting and Control* San Francisco Holden-Day, 1970.

---

[16] The usual caveats associated with most econometric work would include (1) the assumption of structural stability over the sample period, (2) the absence of an "omitted variable" problem where both the rate of inflation and the aggregative policy measures are associated with some third omitted variable The unusual caveats would include (3) the possible sensitivity of the substantive results to alternative filters used to prewhiten the time series (Feige and Pearce 1974), (4) the sensitivity of the results to the seasonal differencing procedures (Pierce 1974), and (5) temporal aggregation problems associated with the use of discrete time series if the "true" model is cast in terms of continuous time (Geweke 1974)

Box, G. E. P., and Newbold, P. "Forecasting Using Leading Indicators " Technical Report no. 251, Dept. Statis., Univ. Wisconsin, 1972.

Cagan, P. "The Monetary Dynamics of Hyperinflation." In *Studies in the Quantity Theory of Money*, edited by M. Friedman. Chicago Univ. Chicago Press, 1956.

Cramer, R. H., and Miller, R. B. "Dynamic Modelling of Multivariate Time Series for Use in Bank Analysis " Mimeographed. Graduate School Bus., Univ. Wisconsin, 1974.

Feige, E. L. "The Consequences of Journal Editorial Policies and a Suggestion for Revision." *J P E* 83, no. 6 (December 1975). 1291-96

Feige, E. L., and Pearce, D K "The Causality Relationship Between Money and Income· A Time Series Approach " Paper presented at the Midwest Economic Association meetings, Chicago, Ill , April 1974

Geweke, J. "Temporal Aggregation in the Multivariate Regression Model " Mimeographed. Minneapolis Univ Minnesota, 1974

Granger, C W J "Investigating Causal Relations by Econometric Models and Cross-spectral Methods " *Econometrica* 37 (July 1969) 424-38

Granger, C. W J , and Newbold, P "Spurious Regressions in Econometrics." *J Econometrics* 2 (July 1974) 111-20

Haugh, L "The Identification of Time Series Interrelationships with Special Reference to Dynamic Regression Models " Ph D dissertation, Univ Wisconsin, Madison, 1972

Hester, D "Stock Prices and the Nixon Speech " SSRI Workshop Series no. 7229, Univ Wisconsin, 1972

Hicks, J R *Value and Capital* Oxford Oxford Univ Press, 1946

Katona, G "Theory of Expectations " In *Human Behavior in Economic Affairs, Essays in Honor of George Katona*, edited by B Strumpel et al San Francisco Jossey-Bass, 1972

Koyck, L M *Distributed Lags and Investment Analysis* Amsterdam North-Holland, 1954

Lovell, M C "Data-Grubbing " Paper presented at the meetings of the Econometric Society, San Francisco, December 1974

Mincer, J "Models of Adaptive Forecasting " In *Economic Forecasts and Expectations*, edited by J Mincer New York Nat Bur Econ. Res , 1969

Modigliani, F , and Shiller, R J "Inflation, Rational Expectations, and the Term Structure of Interest Rates " *Economica* 40 (February 1973) 12-43

Muth, J F "Optimal Properties of Exponentially Weighted Forecasts." *J American Statis Assoc* 55 (June 1960) 299-306

——— "Rational Expectations and the Theory of Price Movements " *Econometrica* 29 (July 1961) 315 35

Naylor, T H , Seaks, T G , and Wichern, D W "Box-Jenkins Methods An Alternative to Econometric Models " *Internat Statis Rev* 40 (August 1972) 123-37

Nelson, C R "The Prediction Performance of the FRB-MIT-Pen Model of the U S Economy " *A E R* 62 (December 1972) 902-18

——— "Rational Expectations and the Predictive Efficiency of Economic Models." *J Bus* 48 (July 1975) 331-43

Niederhoffer, V "The Analysis of World Events and Stock Prices." *J Bus*. 44 (April 1971) 193-219

Pearce, D. K "The Formation and Economic Effects of Price Expectations." Ph.D. dissertation, Univ Wisconsin, Madison, 1974.

Pierce, D. "Relationships—and the Lack Thereof—between Economic Time Series, with Special Reference to Money, Reserves and Interest Rates." Special Studies Paper no 55 (Washington Federal Reserve Board, 1974)

Sargent, T. J "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment." *Brookings Papers on Economic Activity*, no. 2 (1973).

Sargent, T J , and Wallace, N. "Rational Expectations and the Dynamics of Hyperinflation " *Internat. Econ. Rev.* 14 (June 1973). 328–50.

Sims, C. A "Money, Income and Causality " *A E.R* 62 (September 1972): 540–52

U S. Department of Commerce, Office of Business Economics *1971 Business Statistics* Washington Dept Commerce, OBE, 1972.

Walters, A A "Consistent Expectations, Distributed Lags, and the Quantity Theory " *Econ J* 81 (June 1971): 273–81.

Whittle, P *Prediction and Regulation by Linear Least Squares Methods*. Princeton, N J Van Nostrand, 1963

Wold, H O A *A Study in the Analysis of Stationary Time Series* 2d ed Stockholm Almqvist & Wiksell, 1954

Zellner, A , and Palm, F "Time Series Analysis and Simultaneous Equation Econometric Models " *J Econometrics* 2 (May 1974) 17–54

Uncertainty, Waiting Time, and Capacity Utilization: A Stochastic Theory of Product Quality

---

Arthur De Vany

Texas A & M University

The behavior of a monopoly is analyzed in the following setting. Search creates a stochastic stream of customers. Searchers observe the queue and leave if it is too long. Production is random. In the optimal solution, waiting time, effective demand, and the utilization of capacity are jointly determined. Product quality in the form of waiting time is endogenous to the model, and it is found that price or capacity changes affect both the quantity and the quality of output. If an equilibrium exists, there is excess capacity. Price equals marginal cost, but monopoly produces too little capacity and output unless profits are zero.

Queuing is a phenomenon that is largely ignored in economic theory. While queues exist in virtually every type of service process, the standard model of the firm cannot explain their existence without falling back on the idea that the price charged by the firm experiencing the queue is somehow out of equilibrium. Clearly, it is less than satisfactory to resort to a disequilibrium explanation for a phenomenon so persistent and durable as queuing. In this paper an alternative explanation is offered which is capable of explaining the existence of a queue in a process which is in equilibrium. More fundamentally, the paper is an analysis of the demand and supply of waiting time and seeks to explain how the time cost of consumption is determined in market equilibrium under monopoly. The model is based on three propositions: (1) consumer demand is random, (2) production is random, and (3) as a consequence of (1) and (2) the sum of money price and time cost is uncertain, and this uncertainty creates positive net benefits of search for short queues or low time prices.

The valuable comments of Armen Alchian, Tom Saving, Charles Maurice, and Jack Meyer are gratefully acknowledged as well as the excellent criticisms of the anonymous referees. This work was partially supported by the Texas A & M Organized Research Fund.

[*Journal of Political Economy*, 1976, vol. 84, no. 3]

© 1976 by The University of Chicago. All rights reserved.

I argue that a rational consumer adopts a stopping rule in his search for low waiting time He will join only those queues whose length is less than or equal to some critical value The critical queue length depends on the distribution of queue lengths, the value of time, the cost of search, and the availability of alternative suppliers The search procedure and related information structure have an important influence on the degree of monopoly power possessed by the firm

Because customers balk at long queues, the firm will find that both its price and its capacity influence the effective demand for its product A higher price reduces the rate at which customers arrive and also reduces average waiting time Thus, the model suggests that product quality varies inversely with the rate of output when the size of the plant is held constant Greater capacity shifts the effective demand curve to the right This rightward shift of demand also corresponds to the notion of an improvement in product quality because, with more capacity, the plant becomes less congested, waiting time falls, and consumers demand more of the product at each price Thus, the model offers a unique insight into the interrelationship of demand, capacity utilization, and product quality[1]

Demand and capacity utilization are determined by the time value of consumers and their search procedure[2] The steady-state demand function depends on capacity utilization, but utilization also depends on demand This means that demand and capacity utilization, along with the related level of product quality, are mutually determined A necessary condition for the existence of a steady-state demand function in this simultaneous system is that there be excess capacity In exploring the optimal policies of a firm in a setting where "excess" capacity is productive, I obtain a number of new results which have far-reaching implications for the theory of monopoly These results form the basis of a new model of capacity utilization which stresses the value of time and the stochastic nature of search and production One virtue of the focus on waiting time and capacity utilization is that it provides a more or less concrete interpretation for the concept of product quality and overcomes some of the vagueness which has frustrated work in this area Once quality is made endogenous to the theory of monopoly, it becomes clear that price is the marginal revenue of output when product quality is held constant. The marginal revenue schedule of the conventional theory of monopoly, which does not incorporate product quality, is shown in the present model to be the marginal revenue of output when product quality is allowed to vary directly with price This result would appear to dispel some of the traditional objections to monopoly which are based on the divergence of marginal revenue from price While this is partly true, for example, the monopoly outcome is efficient if equilibrium output occurs at the zero profit level, there is still a source of inefficiency present in the model It is shown that monopoly produces too little output The reason for the underproduction of output is that monopoly marginal cost exceeds social marginal cost because the monopolist internalizes the diminishing returns to product quality

---

[1] In the context of regulated markets, a simple version of this model is discussed in De Vany (1975*a*, 1975*b*), with the former containing econometric results There has been a good deal of work done on capacity utilization, which is reviewed by Winston (1974), but this work has not focused on the relationship between capacity utilization and customer waiting time The existence of a relationship between capacity and waiting time is well recognized (Alchian 1970), but this paper is the only attempt to present a reasonably complete model of the problem Other treatments of the queuing problem with balking are those of Noar (1969) and Yechiali (1971) These papers provide an excellent analysis of the optimal congestion toll for a congested service facility with fixed capacity They do not address the problem studied here, which involves the selection of price and capacity under monopoly

[2] The model integrates Stigler's (1961) analysis of search with Becker's (1965) analysis of the allocation of time

I. A Description of the Model

Let a single firm be producing and selling some product or service at a fixed location Orders are filled only through production The firm holds no inventories Demand is less than perfectly elastic in a sense to be made precise below When a potential customer arrives at the facility (a plant or store), he finds the price of the product posted over the order desk There may be a queue of customers at the desk when he arrives, and he notes the posted price and length of the queue and decides whether to join the queue or leave

Now suppose the firm leaves the posted price unchanged for a period of time so that buyers become well informed as to price But let the arrival rate of potential customers be random and also let the rate at which orders may be filled be subject to some random variation Then the expected cost of buying the product or service is estimated by acquiring information through a visit to the facility and observation of the queue at that time If the process is in equilibrium, there will be some distribution of queue lengths from which the searcher calculates an optimal stopping rule Based on some acquired or a priori knowledge of the probability distribution of queue lengths, the searcher chooses a number such that, if the observed queue is less than or equal to this value, he joins it and waits to be served If the line is longer than this reservation length, the searcher balks and refuses to join the line

From the firm's point of view, the search process generates random arrivals of potential customers When the queue length is greater than the critical value, these arrivals refuse to join the queue and depart The queue fluctuates in length, but, if an equilibrium exists, there will be some long-run relative frequency of balks to arrivals and the firm may form an expectation of the number of customers who will demand service over some time interval. The firm can change its posted price, but it must remain at some value long enough for the queuing process to reach equilibrium, if one exists The firm can alter its capacity and thereby change the distribution of queue lengths and the proportion of arrivals who balk This makes effective demand depend on the firm's capacity, and increases in capacity shift the demand function to the right.

Since both service and demand are random, there are times when the service capability is high but there are no customers available to be served To allow for this possibility, the mean output capacity must exceed the mean rate of demand if a steady state is to be achieved So, on average, the firm carries excess capacity in that its mean output capability exceeds its mean demand At times, however, the productive capacity is fully utilized and a queue forms At other times, the plant is idle because there is no one in the queue or being served

When the plant's mean capacity is expanded, the balking value of its customers rises, fewer arrivals balk, and effective demand increases Product quality also improves because the average queue length falls when the rate of utilization of capacity declines Economies of scale exist in the production of waiting time because any queue is more rapidly served for a given ratio of demand to capacity by a bigger plant The basis of natural monopoly is this decreasing cost of product quality

As a physical analogy to the model, one might consider a stream flowing into a pond If we double the mean inflow and outflow of the pond, the average level of the pond will be unaffected, but the average time a unit of water is in the pond will be halved With balking, the incoming stream is diverted whenever the pond reaches a certain level To reach an equilibrium level, the mean inflow, net of diversions, must be less than the mean outflow This is because some of the outflow capacity will occur when the pond is empty and therefore will be ineffective To complete the analogy to the balking model, we must imagine, as will be shown, that the level to which the pond may rise before diversion occurs increases when the outflow capability is increased

II. The Customer Stream

For some class of customers, the firm is the preferred supplier of some product or service Customer preference may be based on locational advantages, lower price, or the fact that the firm is a strict monopoly Even in the last case, however, we allow for the availability of an alternate form of the service in the later development of the model For example, some city may be served by a monopoly airline, but travelers do have the option of travel by other modes Demand is random A stream of customers arrives at the firm, one arrival at a time, but the time between successive arrivals is random[3] I assume that the arrival stream is a stationary stochastic process, with intervals of time between successive arrivals independently distributed There is no periodic, recurrent peaking of demand The firm knows that peaks in demand occur and can even assign a probability to there being n arrivals in some time period It cannot, however, predict when these peaks will occur, since they are completely random These assumptions correspond to postulating that the density function for the interval of time between successive arrivals is a negative exponential distribution[4]

Let price and interarrival times be distributed such that the conditional distribution of interarrival times on price, p, is

$$A(t, p) = \lambda(p)e^{-\lambda(p)t},$$

with mean time between arrivals $= 1/\lambda(p)$ and mean arrival rate $= \lambda(p)$ The arrival rate $\lambda(p)$ is a function of price, with $\lambda_p = d\lambda(p)/dp < 0$ An increase in price moves the distribution of t to the right, causing the probabilities of small interarrival times to fall and those of large interarrival times to rise The rightward shift in the demand distribution corresponds to what we would call a reduction in the quantity demanded, a higher price spreads out the time between customer arrivals[5]

III. The Stochastic Production Function

Assume there is some variation in the time required by the production facility to fill orders[6] Uncertainty in the production process may arise because of random machine failures, lack of critical inputs at unpredictable times, or a variety of other reasons For simplicity I assume production is also a stationary stochastic process so that the output achievable for a given plant or service facility over some time interval is independent of the output achievable in the prior or succeeding time interval. For example, if the relevant time interval is the daily production period of the plant, then the production which could be obtained if the plant produced its maximum output on Tuesday is uncorrelated with the maximum output which could be obtained on Wednesday. This would be the case, for example, if machine failures were random and the failures were repairable overnight.

---

[3] The following draws on basic elements of queuing theory The reader may wish to refer to a standard operations research text to review the theory or to a more thorough treatment such as that of Gnedenko and Kovalenko (1968) or Gross and Harris (1974)

[4] Periodic, regular peaking of demand is ruled out if the renewal distribution is *stationary I am not too comfortable with the temporal independence of the arrival process*, even though there is a fair amount of evidence supporting it If the searcher encounters an excessive queue in the prior period, he may return in the subsequent period, meaning arrivals are not temporarily independent We can eliminate this difficulty by assuming that the service is purchased elsewhere by the balking consumer In addition, many of the results obtained here hold for far more general arrival distributions, so the assumption of independence is not as restrictive as one might think

[5] This is the sense in which demand is assumed to be less than infinitely elastic Were it perfectly elastic, interarrival times would go to infinity as price is increased

[6] Recall that there are no inventories and that production takes place on demand The exponential service time distribution implies that service time on a unit order is independent of service time on preceding orders This assumption fits a large number of service processes, such as communications or haircutting It may not fit production plants so well where successive service times could be correlated In this case, the exponential distribution will be an approximation to the service distribution

Formally, I assume service time s is distributed exponentially:

$$g(s) = \mu e^{-\mu s}, \qquad s \geq 0, \tag{3.1}$$

with mean service time $1/\mu$.

To relate this distribution to the production function usually used by economists, let K and L be the stocks of capital and labor employed by the firm. The "server" in this model is the production line, and μ, the mean output rate when the production line is busy, is a function of capital and labor:

$$\mu = \mu(K, L) \geq 0. \tag{3.2}$$

The production function is increasing and concave in K and L. The conditional distribution of service time s for levels of K and L is obtained by substituting the production function into the service time distribution:

$$S(s, K, L) = \mu(K, L)e^{-\mu(K,L)s}, \qquad s \geq 0, K \geq 0, L \geq 0. \tag{3.3}$$

This distribution is a function of K and L, and the expected marginal products of labor and capital are

$$\frac{\partial S}{\partial L} = e^{-\mu(K,L)s}\mu_K[1 - \mu(K, L)s], \tag{3.4}$$

$$\frac{\partial S}{\partial L} = e^{-(K,L)s}\mu_L[1 - \mu(k, L)s]. \tag{3.5}$$

Since $\partial\mu(K, L)/\partial K = \mu_K > 0$ and $\partial\mu(K, L)/\partial L = \mu_L > 0$, these derivatives indicate that capital and labor augmentation shifts the service time distribution to the left, raising the probability of shorter service times and lowering the probability of longer service times.[7]

There are two notions of capacity implicit in this formulation. I am not defining capacity in the sense often used as being the level of output at which minimum average cost is achieved for a given plant. Nor am I defining capacity as the technical limit on output where the short-run marginal cost of output approaches infinity. There is a maximum production obtainable from the *given* resources of the firm. Whenever the plant is in operation, it produces output at the rate that can be achieved when the given resources of the firm are fully utilized. Since this rate of achievable output is random, it is also appropriate to talk about the mean output of the plant if it were in operation over a suitably long period of time. I define capacity as this quantity. Since from (3.2) the expected service time on a unit order is $1/\mu$, capacity is μ. Also, note that the variance of service time is $1/\mu^2$, therefore, a larger capacity reduces both the mean and the variance of service time. Service becomes quicker and more reliable as capacity is expanded.

---

[7] The relationship is one of first-degree stochastic dominance, with $S(s; K_1, L_1)$ dominating $S(s, K_2, L_2)$ when $K_1 > K_2$ and $L_1 > L_2$.

IV. An Elementary Search Model

Since arrivals and capacity vibrate randomly, there will be intervals of time during which more arrivals appear than can be served, and a queue will form. The consumer is assumed to know the price of the product, but without a visit or inquiry at the firm the waiting time is not known. Thus, the full price, which is the sum of money price and the cost of waiting, is unknown and may only be ascertained through observation of the queue. The consumer's problem is analogous to the price search problem, except that in the present case it is the full price which is a random variable rather than the money price.[8] I derive an optimal stopping rule from a simplified search model. The rule stipulates the maximum length queue which an arriving consumer will join. If the queue exceeds this critical length, he refuses to join it and goes to his next-best alternative source of the service or product. This analysis shows how randomness in waiting time triggers a search for alternatives, and also generates a random switching of consumers from the preferred firm to the alternate during periods of excessive queues. Thus, the monopoly power of the preferred firm is eroded by the information about alternatives, and is eliminated completely at times when the queue exceeds the critical length.

Assume the consumer selects a strategy in the following manner. The product or service is supplied by a monopoly at some price p. Consumer surplus is $R(p)$, with $R_p = dR/dp < 0$. There is some best alternative to the monopolist's product with surplus $R^a(p^a)$ and price p^a. The consumer strictly prefers the monopolist's product to the alternate, that is, for $p = p^a$, $R(p) > R^a(p^a)$, meaning that the demand curve for the product lies above the demand curve for the alternate. On arriving, or inquiring about the line, the consumer weighs two alternatives—to join the queue, or to go to the alternate supplier. The net gain of the first choice is

$$G_{j,n-1} = R(p) - wn/\mu \tag{4.1}$$

---

[8] The problem is similar to the price search problems solved by McCall (1965) and by Telser (1973). Here the buyer is assumed to learn the price on the first visit. The price remains fixed, but the queue fluctuates according to a distribution which may or may not be known by the buyer. A trial is a visit to the facility at a different point in time, and a "price" is obtained through observation of the queue. Noar (1969) and Yechiali (1971) solve the optimal balking problem in a setting somewhat more restrictive than the present one.

when there are $n - 1$ customers in the system ahead of the arrival. The expected cost of waiting is wn/μ, with $w =$ the unit cost of time and $n/\mu =$ the conditional expected wait when the arrival is the nth member in the system.

If the customer does not join, he goes to the alternative supplier. Assume that the cost of going from the preferred supplier to the alternate supplier is composed of a time cost wt and a money cost c. Also, assume that the waiting time at the alternate supplier is uncorrelated with waiting time at the preferred supplier and that the steady-state expected number in the alternate supplier's system is L_A Then the expected gain of balking from the preferred supplier to the alternate is

$$G_A = R^a(p^a) - c - wt - w(L^a + 1)1/\mu^a, \qquad (4\ 2)$$

where $1/\mu^a$ is the expected service time of the alternate supplier Assume $G_a > 0$ and that the expected benefits of buying from the preferred supplier are positive, that is, $E(G) = R(p) - wL1/\mu > 0$, where $L =$ expected number in the system of the preferred supplier

An optimal strategy for joining the queue at the preferred supplier is a number n_s which satisfies the inequality

$$R(p) - wn_s1/\mu \geq R^a(p^a) - L - wt - w(L^a + 1)1/\mu^a \qquad (4\ 3)$$

We can solve this inequality for the critical value n_s

$$n_s \leq [R(p) - R^a(p^a) + c]\mu/w + \mu t + (L^a + 1)\mu/\mu^a, \qquad (4\ 4)$$

which is the largest number in the preferred supplier's system such that a customer will enter it Since it will be more convenient in the subsequent analysis to deal with the number at which the customer will balk, we define the balking value as

$$B = [n_s + 1], \qquad (4\ 5)$$

where the bracket notation signifies that B is the largest integer not exceeding $n_s + 1$

The optimal balking function is not differentiable, but the dependence of B on the parameters may be determined through numerical analysis The plus or minus sign above the parameter in equation (4 6) indicates how the balking value changes when a parameter changes in value

$$B = B(\overset{-}{p}, \overset{+}{p_a}, \overset{+}{c}, \overset{+}{t}, \overset{-}{w}, \overset{+}{\mu}, \overset{-}{\mu_a}, \overset{+}{L^a}) \qquad (4\ 6)$$

We can summarize the argument to this point in the following way Customers arriving at the firm will join the queue whenever its length is less than the expected length, they will also join the queue when it is larger than its expected length if the cost of waiting in it is less than the surplus lost plus the expected waiting and search cost of going to the alternate supplier The customer will tolerate longer queues (have a higher balking value) the lower the price of the good, the higher the price of the alternate, the greater the service rate of the firm relative to the service rate of the alternate, the greater the cost of switching to the alternate, and the lower the cost of time

V. The Queue: Effective Demand and Capacity Utilization in the Steady State

We now seek to derive the effective demand function under the following assumptions· (1) The arrival stream is stationary and Poisson with parameter $\lambda(p)$ (2) Customer orders are serviced first come, first served by a plant with exponentially distributed service times with mean $= 1/\mu$ (3) Customers have a common balking value B which depends on the expected benefits of purchase at the firm relative to the expected benefits of purchase from an alternate supplier

As a consequence of these assumptions there is an absolute limit B on the number of customers in the system Our problem then becomes a well-known queuing model, the solution of which we shall merely summarize here [9]

Define

$$\rho \equiv \lambda/\mu \qquad (5\ 1)$$

as a measure of traffic intensity This is a dimensionless quantity which measures the load on the service mechanism or plant The higher ρ is, the greater is the mean rate at which orders arrive relative to the plant's mean capacity to fill them Presently, there are no limits on ρ, the mean arrival rate may be less than mean capacity, or the facility may be overloaded with mean arrivals exceeding mean capacity

Now assume that the process reaches a steady-state equilibrium Haight (1957) has shown that, if an equilibrium exists with $p \neq 1$, the following are true

$$p(x) = \begin{cases} \dfrac{1-\rho}{1-\rho^{B+1}}\rho^x & 0 \le x \le B \\ 0 & B < x, \end{cases} \qquad (5\ 2)$$

$$m = \frac{\rho}{1-\rho} - (B+1)\frac{\rho^{B+1}}{1-\rho^{B+1}}, \qquad (5\ 3)$$

where $p(x) \equiv$ probability there are x in the system and $m \equiv$ mean number in system It is known that an equilibrium exists if $B < \infty$, which is ensured because consumer surplus $R(p)$ is finite

---

[9] The model is the $(M/M/1/B)$ model which denotes Poisson arrivals and service with single server and limit B on system size The study by Gross and Harris (1974) contains a thorough treatment of this model Haight (1957) solves the more general problem of balking when there is a distribution of balking values rather than a single deterministic value His model would seem to be very useful when customers differ in their time values, or when their time values are random but centered around a value such as their wage rate

The probability that someone balks is the probability that the number in the system equals the balking value B. From (5.2) the balking probability is

$$\beta = p(B) = \frac{1 - \rho}{1 - \rho^{B+1}} \rho^{B} \tag{5.4}$$

We can interpret β as the long-run proportion of arrivals who balk. It follows, therefore, that the proportion of searchers who join the queue and proceed to purchase the service is $1 - \beta$. Mean effective demand for the service, therefore, is

$$\lambda' = \lambda(1 - \beta) \tag{5.5}$$

and the load imposed on the system by customers is

$$\rho' = \rho(1 - \beta) \tag{5.6}$$

For finite B the balking probability β is positive so that $\lambda' < \lambda$. In other words, there is a kind of excess demand for the firm's product in the sense that some potential customers arrive but turn away to the alternate because the queue they encounter is at the balking length. These disappointed, would-be buyers can only be eliminated at the limit as the plant's mean capacity is increased indefinitely ($\mu \rightarrow \infty$).

Since the probability of an idle plant is $p(0) = 1 - \rho'$, and $p(0) > 0$, it follows that $\rho' < 1$. This interesting result indicates that the firm's mean output rate is less than the mean production capacity of its plant, in other words, the process will have excess capacity if it achieves a steady-state equilibrium. This result holds even if the arrival rate overloads the capabilities of the plant, that is, even if $\rho > 1$, we will still have $\rho' < 1$. The parameter ρ' can be interpreted either as the probability that the plant is busy or as the fraction of each time interval the plant is busy.

We are now in a position to construct the demand function relating output to price, the balking value, and the level of capacity. Initially, consumers desire to order an average of $\lambda(p)$ units per period. However, given the balking value B, only $\lambda' = \lambda(p)(1 - \beta)$ units are actually ordered. The firm loses $\lambda(p)\beta$ orders each period because customers refuse to join the line when its length exceeds B units. Lost orders, or balks, depend on queue length, which is a function of ρ, and B. Substituting into (5.5) for β from (5.4), and using $\rho = \lambda(p)/\mu$, yields

$$\lambda' = \lambda(p) \left\{ 1 - \frac{1 - [\lambda(p)/\mu]}{1 - [\lambda(p)/\mu]^{B+1}} \left[\frac{\lambda(p)}{\mu} \right]^{B} \right\} \tag{5.7}$$

Noting the dependence of β on p, B, and ρ, we can simplify (5.7) to

$$\lambda' = \lambda'(p, \mu, B) = \lambda(p)\{1 - \beta[\lambda(p), \mu, B]\}. \tag{5.8}$$

In this formulation the equilibrium mean rate of demand is a function of price, mean capacity, and the balking value Recall that we are evaluating the function over a restricted domain, because an equilibrium does not exist where $\rho' \geq 1$ Recall also that $\lambda(p)$ is the mean rate of arrival of customers who are willing to buy at the price p, provided they encounter fewer than B customers, whereas $\lambda'(p, \mu, B)$ is the equilibrium mean number of customers who do buy For this reason I refer to $\lambda'(\cdot)$ as the effective demand function while $\lambda(\ )$ is the potential demand function

As a simplification, I treat the effective demand function as being continuous and differentiable within the interior of the region defined by $\rho'(p, \mu, B) < 1$ [10] The derivatives are

$$\frac{\partial \lambda'}{\partial p} = \lambda'_p = (1 - \beta)\lambda_p - \lambda\beta_p < 0, \tag{5 9}$$

$$\frac{\partial \lambda'}{\partial \mu} = \lambda'_\mu = -\lambda\beta_\mu > 0 \tag{5 10}$$

The first equation indicates that the price affects both the arrival rate of searchers and the proportion of those searchers who stay A higher price reduces both the arrival rate and the balking value, thereby causing a greater proportion of arrivals to balk Consumers arrive less frequently and become more critical of queues when they are paying a high price

The next equation is the derivative of effective demand with respect to capacity Since larger capacity raises the balking value, it follows that $\beta_\mu < 0$ Fewer arrivals balk when μ is increased, because the expected wait associated with any queue length is less when the plant has higher capacity Output increases with capacity because arrivals encounter shorter and more rapidly moving queues

The demand function has a simple interpretation Effective demand lies beneath $\lambda(p)$ because some arrivals balk The effective demand curve is less elastic than the potential curve [11] When the firm adds capacity, the

---

[10] Since B is an integer, it is not a continuous function of price However, for fixed B, the balking probability β is a continuous function of price The result is that β is piecewise continuous with "small" jumps at the points of discontinuity where B takes on a new integer value We lose no insight by approximating β by a continuous function, and make a considerable gain through being allowed to use the calculus and the marginal revenue–marginal cost interpretations that can be placed on the derivatives

[11] This result is easily shown Price elasticity of demand when quality (waiting time) is held constant is $\zeta = \lambda_p(p/\lambda)$ The revealed price elasticity of demand is

$$\zeta' = \lambda'_p \frac{p}{\lambda'} = \lambda_p \frac{p}{\lambda} - \beta_p \frac{p}{(1 - \beta)} = \zeta - \beta_p \frac{p}{(1 - \beta)},$$

which is the true, constant-quality price elasticity, minus a bias term Since $\beta_p < 0$, a positive bias term is added to the true price elasticity, making the absolute value of the revealed elasticity ζ' less than the true elasticity ζ The more congested the facility, the closer is β to 1 and the greater is the bias This finding would seem to be useful in demand studies of service industries and communications where congestion influences consumer demand De Vany (1975*b*) finds that the constant-quality price elasticity of demand for air travel is larger than the revealed price elasticity

$\lambda'(p, \mu, B)$ curve is shifted to the right, but it only approaches the $\lambda(p)$ curve as μ approaches infinity

VI. Market Equilibrium

Let the expected cost function be defined on all λ' and μ such that $\lambda' < \mu$ and denote the function

$$c = c(\lambda', \mu). \qquad (6\ 1)$$

The derivative $c_{\lambda'} > 0$ is the marginal cost of mean output where the variation in λ' is subject to $\lambda' < \mu$ The derivative $c_\mu > 0$ is the marginal cost of a higher level of planned mean capacity

Assume that the firm's objective is to maximize expected profits The firm's attitude toward risk is ignored here

The expression for expected profits is

$$E(\pi) = p\lambda'(p, \mu, B) - c[\lambda'(p, \mu, B), \mu] \qquad (6\ 2)$$

Since β is a monotone decreasing function of μ, revenue is a concave function of μ Revenue is also concave in price Costs may be concave or convex, but the expected profit expression is assumed to be concave in p and μ The Kuhn-Tucker conditions are

$$p \quad (p - c_{\lambda'})[(1 - \beta)\lambda_p - \lambda(p)\beta_p] + \lambda(p)(1 - \beta) \leq 0 \qquad (6\ 3)$$

$$\mu \quad -(p - c_{\lambda'})\lambda(p)\beta_\mu - c_\mu \leq 0 \qquad (6\ 4)$$

The first term of (6 3), $p - c_{\lambda'}$, is the expression for expected net short-run revenue, that is, net revenue when expected output is varied within the limits of the capacity constraint Clearly, this term must be positive, so that price exceeds the short-run expected marginal cost of output [12]

The bracketed term in (6 3) is the change in output induced by a price change (see [5 9] above) The product of this term with $p - c_{\lambda'}$ is the marginal revenue of a price change holding $p - c_{\lambda'}$ unchanged The term $\lambda(p)(1 - \beta)$ is the marginal revenue of a price change holding arrivals unchanged

Equation (6 4) indicates that capacity is expanded to the point where the net marginal revenue achieved through reduction of the searchers who balk equals the marginal cost of capacity

It is useful to think of (6 3) and (6 4) as short- and long-run conditions Equation (6 3) is a short-run condition in the sense that the size of the plant is fixed and output changes are accomplished by changing price. This short-run condition may be written as

---

[12] Stigler (1961) obtains a similar result in analyzing advertising, though his setting is one of certainty for the firm. This theorem is analogous to the traditional monopoly result that $p > mc$, except that we have shown it to be a short-run condition

$$p(1 + 1/\varepsilon) = c_{\lambda'} \tag{6 5}$$

where ε is the price elasticity of effective demand $\lambda'_p p/\lambda'$. This is a traditional monopoly result in that marginal revenue equals marginal cost and price is set in the elastic portion of the demand curve It is important to note, however, that the relevant marginal cost in the present model is strictly a short-run cost, since $c_{\lambda'}$ is the increase in cost leaving capacity fixed Essentially, these costs are associated with the materials consumed in production, such as fuel, steel, electrical energy, and so on

Equation (6 4), which I have identified as a long-run condition, may be written as

$$p = c_\lambda + \frac{c_\mu}{\lambda'_\mu} \tag{6 6}$$

This startling result says that price equals the long-run marginal cost of output To show that the right-hand side of (6 6) is the long-run marginal cost of output, rewrite it as

$$\frac{c_\lambda \lambda'_\mu + c_\mu}{\lambda'_\mu}$$

The numerator is the full incremental cost of increasing capacity This cost consists of the short-run marginal cost of output times the increase in output induced by greater capacity, plus the marginal cost of capacity This full incremental cost of capacity is multiplied by $1/\lambda'_\mu$, the increment to capacity required to induce a unit increase in output Thus, the right-hand side of (6 6) is the full marginal cost of output when output is increased through an expansion of capacity

In the standard model of monopoly, there is no such distinction between short- and long-run marginal revenue We can identify the traditional marginal revenue curve within the present model as the curve indicating the additional revenue coming from an extra unit of output when price is lowered sufficiently to compensate for the additional waiting time imposed on the other users of the system In other words, the traditional marginal revenue appears here as the revenue associated with a unit increase in output allowing product quality to decline. Price is the true marginal revenue of output in the present model, because it is marginal revenue when product quality is held constant with capacity adjusted to leave waiting time unaffected by the increase in output In this sense, the distinction I have made between short- and long-run effects is really a distinction between changes that permit product quality to vary and those which hold it constant

This analysis sheds some light on a long-misunderstood problem in economics It has been said that, if there are empty seats on a train, then the marginal cost of hauling another passenger is zero, and efficiency requires that the price be zero This argument misses the mark on two points First, in the arrival and service process of the queuing model, no passenger is ever left at the platform by a train having empty seats. Every passenger is boarded On average, the train has empty seats because that is a necessary condition for equilibrium This means that it is the train, and the passengers on board, which would be made to wait until the arrival of an additional passenger if the train were not to leave with an empty seat Second, the marginal cost of hauling another passenger is not zero, it is the cost of the delay imposed on those aboard while they wait for the next passenger to arrive If there are E empty seats, passengers already on board will have an expected wait of E/λ before the train departs Another approach is to realize that when we speak of an additional passenger we mean a unit increase in the long-run output rate of the railroad This additional passenger must pay the full marginal cost of his travel The full marginal cost is $c_{\lambda'}$, here equal to zero, plus cost of the incremental capacity required to prevent his usage of the system from imposing delay costs on other users The efficient price is, therefore, $p = c'_\lambda + c_\mu \mu/\lambda'$ rather than $p = c_{\lambda'}$

It trivially follows that the long-run average cost of output is rising in the neighborhood of equilibrium Price equals marginal cost, and price must equal or exceed average cost for the firm to operate, therefore, marginal cost must equal or exceed average cost, and average cost is either rising or at a minimum at equilibrium If profits are zero, average cost is minimized There is no tangency solution on the downward-sloping portion of the average cost curve[13]

VII. Efficiency Considerations

It has been shown that price equals marginal cost, which suggests that the monopoly outcome is efficient In this section it is shown (1) that monopoly produces the efficient level of capacity for any level of output but (2) that monopoly marginal cost exceeds marginal social cost so that output is inefficiently small

First, consider what is required for efficient capacity at any level of output The marginal value of capacity must equal its marginal cost That this is true of monopoly can be shown by solving (6 3) and (6 4) for p, equating and obtaining

$$-\frac{\lambda^{\lambda'_\mu}}{\lambda'_p} = c_\mu \qquad (7\ 1)$$

---

[13] This result echoes the analysis of monopolistic competition of Demsetz (1964)

The term on the left of (7.1) is the rate at which each consumer is willing to trade money for capacity times the number of consumers, and is the marginal value of capacity in the market Thus, the marginal value of capacity equals its marginal cost Note that monopoly marginal cost of capacity is the social marginal cost

The next result to be shown is that monopoly output is underproduced. Since price equals marginal cost, it suffices to show that monopoly marginal cost is greater than social marginal cost Monopoly marginal cost is $c_{\lambda'} + c_\mu/\lambda'_\mu$ Social marginal cost, the forgone value of resources used by the monopolist, is given by

$$MC^* = c_{\lambda'} + c_\mu(\mu/\lambda')$$

The quantity MC^* is the short-run marginal cost of output plus the marginal cost of capacity times capacity per unit of output The concavity of demand in μ implies that $1/\lambda'_\mu > \mu/\lambda'$, hence it follows that $MC > MC^*$

Since the monopolist internalizes the decreasing returns to capacity, his marginal cost of output diverges from social marginal cost This result is analogous to the departure of his marginal revenue from marginal social value in the standard model

VIII. Stochastic Determinants of Natural Monopoly

In this section I explore the sources of decreasing cost that give rise to natural monopoly The point I establish is that the source of decreasing cost relates to product quality, not output

Consider the equilibrium waiting time obtained from Little's formula

$$W = L/\lambda', \tag{8.1}$$

where L = expected number in the system Since $\lambda' = \lambda(1 - \beta)$, we have

$$W = L/\lambda(1 - \beta) \tag{8 2}$$

Suppose λ and μ are values such that $\rho = \lambda/\mu \neq 1$ The value of ρ, along with B, completely determines both L and β

Now let the size of the market and scale of the plant double; this amounts to a doubling of output and capacity while the ratio of arrivals to capacity remains unchanged What happens to expected wait and the rate of utilization of capacity? If there were no balking, a doubling of λ and μ would leave ρ and L unchanged and would, therefore, cause W to fall to half its initial value, as is evident from equation (8 2) when $\beta = 0$ Capacity utilization is unchanged This means a large firm has a lower waiting time than a small firm when both utilize their capacity at equal intensity, that is, there are economies of scale in the production of quality or waiting time

When there is balking, a similar result holds but capacity utilization increases with scale. A doubling of scale (in the constant ρ sense) raises the balking value (see eq. [4.6]). This decreases β, so λ' rises by more than λ, and there is a "small" increase in L. This means $\lambda' = \lambda(1 - \beta)$ rises by more than L when λ and μ are doubled, and W falls. That is, with balking, a proportional increase in the size of the market and plant reduces expected wait and increases the effective utilization rate of the plant.

The kind of scale expansion I have described is ordinarily not available to a firm, because it requires a rightward shift of the demand function $\lambda(p)$. The firm may only influence effective demand $\lambda'(p, \mu, B)$ through choice of p and μ. For given p, higher μ shifts effective demand to the right, but by decreasing amounts. Thus, at a given price, we expect decreasing returns to capacity, with ρ and ρ' falling as μ is increased.[14] The explanation for this result is simple. When the size of the market is fixed, expansion of capacity is tantamount to quality improvement, and there are diminishing returns on the demand side to product quality.

As an application of these findings, consider the utilization rate of hospital beds. Assume that the rate of arrivals is proportional to the number of beds in each hospital, so that ρ is the same for both. If there were no balking, then each hospital would have the same occupancy rate, but the larger hospital would have a shorter average wait for a room because of the economies of scale for waiting time. We do expect balking, however, since prospective patients have high time value. Since the balking value depends only on the capacity of the hospitals, assuming their room rate is the same, the balking value for the larger hospital will be greater. But, because of economies of scale, the number of patients waiting for a room will reach the balking value less frequently in the large hospital than in the small hospital, so that a smaller proportion of arrivals balk at the large hospital. For this reason, the larger hospital has a larger ratio of effective arrivals to beds than has the smaller hospital. Its occupancy rate is greater, and it captures a share of the market which is greater than its share of beds. This result is quite general and seems to explain why the share of the market captured by large airlines is greater than their share of flights (Douglas and Miller 1973), or why electrical utilities in large market areas operate at a higher rate of utilization of capacity than do utilities in small market areas (Smith 1975).

---

[14] One way to test whether capacity has any value to consumers is to observe the change in the utilization of capacity when capacity is increased. Where capacity has no value, $d\rho/d\mu = -\lambda/\mu^2$. Where capacity does have value,

$$\frac{d\rho'}{d\mu} = (\beta - 1)\frac{\lambda\rho}{\mu^2} - \beta_\mu \frac{\lambda}{\mu}$$

Since $\beta > 0$ and $\beta_\mu < 0$, we have $d\rho/d\mu < d\rho'/d\mu < 0$.

IX. Further Issues and Observations

The view of this paper is that, when consumers are impatient and there is a stochastic element in their demand behavior, the capacity of the firm becomes an argument in the effective demand function I have shown how to obtain the effective demand function from a search process through the use of queuing theory, and have further shown how equilibrium price, output, and capacity are determined

One result that has important extensions is the finding that the monopolist's price contains a charge for the short-run marginal cost of output and an additional charge for the capacity provided which takes the form of a congestion toll The congestion toll is not the result of periodic peaking of demand, since it is assumed that a stationary distribution generates arrivals Incorporation of periodic demand into the present model would appear to give a more realistic picture of the peak load problem than is given by the usual model, which implicitly assumes that the quality of the service has no effect on demand [15]

The firm's attitudes toward risk have been ignored here by assuming maximization of expected profits It is shown that raising price or the level of capacity reduces the variance of the arrival distribution This suggests that a risk-averse firm might choose to sacrifice some expected return for a smaller variance by charging a higher price and providing more capacity than a risk-neutral firm On the other hand, the firm's risk aversion may match up with the risk aversion of its consumers, for example, they both wish to avoid overloading of the plant The model permits the risk preferences on both sides of the market to be studied

Further exploration is needed in order to assess the results of the usual measures employed in dealing with monopoly What are the consequences of the regulated price for the quantity and quality of output? How does rate-of-return regulation affect the outcome? Another set of issues relates to the transient behavior of the system Is there a unique path from one equilibrium to another? Does the queuing process break down along some paths so that a new, feasible equilibrium is never reached? It is reasonable to suppose that work in this area would contribute to a theory of bankruptcy and enlighten understanding of capacity utilization over business fluctuations

We have not here allowed for the establishment of priority queues, service has been provided on a first come, first served basis One could study the process when one or more queues are established with different priorities of service, a case which would arise when consumers have different time values In addition, one would like to study the trade-offs between price variability and variability in waiting time My speculation is that in many situations it is cheaper for the consumer to carry an inventory of time than it is to carry the higher cash balances that would be required if prices were continuously varied to keep waiting time at some low, constant value

---

[15] That reliability does influence the demand for electricity is shown by Smith (1975) Smith's dissertation is a test of an adaptation of the model developed here

Aspects of industry structure might be usefully analyzed with the model It would seem, for example, that the alleged price rigidity of oligopoly might be explained in terms of the pooling of demand variances in a few large firms, which would have a dampening effect on price fluctuations Under competition, each individual firm experiences more variability and may adjust price more often On the other hand, consumer search tends to even out the demand fluctuations experienced by individual firms If the cost of search is low enough, it may be possible for a competitive market to act as one big firm in allocating customers to stores which are not busy If this were possible, competition might internalize the economies of scale and produce the efficient levels of waiting time and output (De Vany and Saving 1975)

References

Alchian, A A "Information Costs, Pricing, and Resource Unemployment" In *Microeconomic Foundations of Employment and Inflation Theory*, edited by E S Phelps et al New York Wiley, 1970

Becker, G S "A Theory of the Allocation of Time" *Econ J* 75 (September 1965) 493-517

Demsetz, H "The Welfare and Empirical Implications of Monopolistic Competition" *Econ J* 74 (September 1964) 623-41

De Vany, A S "Capacity Utilization under Alternative Regulatory Restraints An Analysis of Taxi Markets" *J P E* 83, no 1 (February 1975) 83-94 (*a*)

——— "The Effect of Price and Entry Regulation on Airline Output, Capacity and Efficiency" *Bell J Econ and Management Sc* 6 (Spring 1975) 327-45 (*b*)

De Vany, A S, and Saving, T R "Truck Transportation Efficiency" Report prepared for the Motor Vehicle Manufacturers Association, Summer 1975

Douglas, G W, and Miller, J C, III *Economic Regulation of Domestic Air Transport Theory and Policy* Washington Brookings Inst, 1973

Gnedenko, B V, and Kovalenko, I N *Introduction to Queueing Theory* Jerusalem Israel Program Sci Translations, 1968

Gross, Donald, and Harris, Carl M *Fundamentals of Queueing Theory* New York Wiley, 1974

Haight, F M "Queueing with Balking" *Biometrika* 44 (December 1957) 360-69

McCall, J J "The Economics of Information and Optimal Stopping Rules" *J Bus* 38 (July 1965) 300-317

Noar, P "The Regulation of Queue Size by Levying Tolls" *Econometrica* 37, no 1 (January 1969) 15-23

Smith, E "Reliability in the Demand for and Supply of Electricity" Ph.D dissertation, Texas A & M Univ, 1975

Stigler, G "The Economics of Information" *J.P E* 69, no. 3 (June 1961) 213–25
Telser, L G. "Searching for the Lowest Price" *A E R.* 63, no. 2 (May 1973) 40–49
Winston, G C. "The Theory of Capital Utilization and Idleness" *J. Econ Literature* 4 (December 1974). 1301–20
Yechiali, U "On Optimal Balking Rules and Toll Charges in the *GI/M/1* Queueing Process" *Operations Res* 19, no 2 (1971) 349–70

Project Evaluation, Shadow Prices, and Trade Policy

Ronald Findlay and Stanislaw Wellisz

Columbia University

The problem of how to determine the appropriate shadow prices of primary inputs for the evaluation of new projects in an open economy subject to distortions is discussed These shadow prices are compared with the corresponding free-trade and actual market prices It is shown that if the distortion is an output subsidy or tax on existing production, the optimal intervention for new projects is subsidies and taxes on primary factors equal to the difference between the shadow prices and the market prices and not an output subsidy or single shadow exchange rate to provide offsetting protection for the new project It is also shown that projects viable under free trade may reduce welfare if they are introduced into a distorted economy, while projects that would increase welfare in these circumstances might not be viable under free trade

The general concern with the problems of accelerating the growth of the less developed countries has led to an extensive technical literature on the appropriate concepts and methods to use in the evaluation of development projects Since the price systems of most less developed countries are notoriously subject to distortions of every conceivable variety, a major issue in this field has been the specification of a set of "shadow prices" that would truly reflect the relevant costs and returns of the inputs and outputs involved in a project

In the fifties and sixties the fashionable approach to this problem was in terms of multisectoral programming models, of which the contributions by Hollis Chenery (1961) and Michael Bruno (1967) are perhaps the best known More recently, an OECD manual on methods of cost-benefit analysis for developing countries by Little and Mirrlees (1969) has

We are grateful to G Calvo, T McCool, C Rodriguez, and an anonymous referee for valuable comments We alone, of course, bear responsibility for all errors

[*Journal of Political Economy* 1976, vol 84 no 3]

© 1976 by The University of Chicago All rights reserved

advocated the use of "world prices" of traded goods as the relevant shadow prices, thus apparently cutting through a maze of heavy computational requirements They elaborate and clarify the approach of the OECD manual in Little and Mirrlees (1974).

The Little-Mirrlees method has given rise to controversy and misunderstanding in spite of, or perhaps because of, its widespread adoption by practitioners This has led to several attempts at exposition and clarification of the logical basis of their method, of which the most helpful known to us have been by their Oxford colleagues Joshi (1972), Lal (1974), and particularly Corden (1974) In this paper we follow him in looking at the problem of project evaluation in terms of the "second-best" welfare analysis of commercial policy and the theory of optimal intervention in a distorted open economy, using the familiar two-good, two-factor model to examine the relationship between market prices, free-trade prices, and second-best shadow prices of primary factors Corden confines his analysis to simple numerical examples, as does Lal At the other extreme is a very elaborate and mathematically complex model devised by Dasgupta and Stiglitz (1974) Insofar as we can understand them, their emphases, concerns, and results appear to be somewhat different from those of Corden and ourselves

The rest of this paper is in two parts The first examines the Little-Mirrlees second-best shadow factor prices and the nature of the optimal intervention required to implement project evaluation based on it The second part compares this approach to the proposal by Bacha and Taylor (1971) that projects be selected on the basis of whether or not they would be viable under free trade

I

Consider the standard perfectly competitive two-factor two-good model of trade theory Two inelastically supplied factors called labor and "capital" produce two goods under conditions represented by constant returns to scale production functions that permit smooth substitution between the two inputs Suppose, further, though it is not really necessary, that the strong factor intensity condition is satisfied so that one of the goods is always more capital intensive at any set of factor prices Then, given relative product prices, relative factor prices are determined

If the country is "small," so that it faces exogenously given world prices for the products, the value of production and consumption are both maximized at the given world prices under free trade Let us now suppose that trade is restricted by a subsidy on the production of the importable, a tax on the production of the exportable, or both in combination These subsidies and taxes are purely distortional since, by hypothesis, there is no corrective function for them to fulfill. Relative product prices to domestic firms are now altered, and by the Stolper-Samuelson theorem the price of the factor used intensively in the importable sector will be raised and the other one will be lowered. In our case, let us assume that the importable is capital intensive so that the restrictive trade policy raises the rental on capital and reduces the real wage. As a result of the distortion, production is no longer maximized at world prices. Trade still takes place at world prices and consumers equate marginal rates of substitution to world prices, since we have assumed that protection takes the form of a subsidy on the output of the importable or a tax on the output of the exportable rather than an import tariff, which would impose an additional consumption cost.

We now introduce the possibility of producing a third good, which also has a fixed price on the world market and a constant returns to scale neoclassical production function, as a new project into the distorted economy. The consumption aspects of the good are irrelevant since with fixed world prices production and consumption can be treated separately. Welfare in terms of whatever basket of goods it is desired to consume (or invest) at the fixed world prices would clearly increase if introduction of the new good resulted in an increase in the value of domestic production at world prices, since this would increase the purchasing power of the nation in the world market. Hence, the criterion for acceptance of the project is whether or not it will increase the value of total production at world prices as compared with the existing situation, assuming that the distortional policy on the existing goods continues unchanged. This makes the problem an inherently second-best one.

Suppose that world prices are measured in terms of some numeraire such as the U.S. dollar. The return from producing one unit of the new good would be the fixed price of this good on the world market. What would be the cost? If a unit of capital or labor is released from existing production there will be some reduction in the value of this production at world prices. This would measure the appropriate shadow price in each case, according to the Little-Mirrlees method. Using these shadow prices, we can find the combination of capital and labor that minimizes the cost of producing one unit of the new commodity at these prices. If the total cost in terms of shadow prices of this efficient combination is greater than the world price, then the project should be rejected. If it is less it should be accepted, and if it is equal the project would be exactly on the margin.

Formally, the situation can be analyzed by some elementary notions of linear programming. In terms of world prices, the initial situation can be described as the solution to the following problem: maximize $V = p_1^*X_1 + p_2^*X_2$ subject to $a_{11}^0X_1 + a_{12}^0X_2 \leqq L$, $a_{21}^0X_1 + a_{22}^0X_2 \leqq K$, where p_1^* and p_2^* are the world prices, X_1 and X_2 are the output levels of the two goods in the restricted trade situation, and the a_{ij}^0 refer to the capital and labor coefficients per unit of each output corresponding to the techniques chosen by domestic producers in that situation

These technical coefficients are determined by the behavior of domestic producers who are interested in maximizing profits at the tax and subsidy inclusive prices Their problem is to maximize $Z = (1 + s)p_1^*X_1 + (1 - t)p_2^*X_2$ subject to $X_1 = X_1(L_1, K_1)$, $X_2 = X_2(L_2, K_2)$, $L_1 + L_2 \leqq L$, $K_1 + K_2 \leqq K$, where s and t are the subsidy and tax, respectively, either but not both of which may be zero.

The solution to this nonlinear problem gives the a_{ij}^0 of the first problem on division of X_1 and X_2 by L_1, L_2, K_1, and K_2 Given this set of a_{ij}^0, the solution X_1, X_2 that maximizes V is the same as the one that maximizes Z with strict equality in the constraints in both cases

The planner, who is interested in V, uses the socially correct world prices to evaluate X_1 and X_2 but is forced to recognize that resources are used inefficiently by the private sector as the result of the distortion in relative product prices, since by hypothesis there is nothing that he can do to alter the a_{ij}^0's that result from profit maximization by the private sector at those distorted prices

The appropriate second-best shadow prices for the planners to use is given by the dual to maximizing V, which is minimize $W^0 = w_1^0L + w_2^0K$ subject to $a_{11}^0w_1^0 + a_{21}^0w_2^0 \geqq p_1^*$, $a_{12}^0w_1^0 + a_{22}^0w_2^0 \geqq p_2^*$, where w_1^0 and w_2^0 are the shadow prices of labor and capital, respectively With both goods produced in positive amounts in the primal problem, both constraints will hold with equality in the dual yielding w_1^0 and w_2^0 as the solution to the system of two linear equations The decision on whether or not to accept or reject the new project therefore depends upon whether $p_3^* \gtreqless w_1^0a_{13}^0 + w_2^0a_{23}^0$, where a_{13}^0 and a_{23}^0 are the labor and capital coefficients for the third good that minimize its cost of production at the shadow prices w_1^0 and w_2^0, and p_3^* is the fixed world price of the new good The criterion can be written equivalently as $p_3^* \gtreqless p_1^*(\Delta X_1)^0 + p_2^*(\Delta X_2)^0$, where $(\Delta X_1)^0$ and $(\Delta X_2)^0$ are the changes in the levels of X_1 and X_2 that are necessary to release the amounts of labor and capital a_{13}^0 and a_{23}^0 that are needed to produce one unit of the new good

The market prices of labor and capital, however, are the values $\bar{w}_1$ and $\bar{w}_2$ that satisfy $a_{11}^0\bar{w}_1 + a_{21}^0\bar{w}_2 = (1 + s)p_1^*$, $a_{12}^0\bar{w}_1 + a_{22}^0\bar{w}_2 = (1 - t)p_2^*$

It is readily seen that $\bar{w}_2 > w_2^0$ and $\bar{w}_1 < w_1^0$ since protection of the capital intensive good raises the rental on capital and reduces the real wage of labor Consequently, the optimal intervention required to offset the distortion is to give a subsidy of $(\bar{w}_2 - w_2^0)$ on the use of capital and to place a tax of $(\bar{w}_1^0 - \bar{w}_1)$ on the use of labor Producers should receive the world price of the output of the new good—that is, there should be no subsidies on output

In his discussion of the problem, Corden (1974, p 398) states that projects accepted on the Little-Mirrlees criterion would require an output subsidy to make them viable in the distorted market situation He appears here to be implicitly assuming fixed coefficients for the new good, in which case he would be correct The distortion of relative factor prices is of no consequence when there is no choice of technique problem Any method that makes domestic producers break even is acceptable, and, in terms of our notation, an output subsidy on the new good of $q = \tilde{w}_1 a^0_{13} + \tilde{w}_2 a^0_{23} - p^*_3$ would be equivalent to our capital subsidy and labor tax. With variable coefficients, however, an output subsidy alone would never offset the distortion in relative factor prices, and so it cannot be the optimal method of intervention It also follows for the same reason that a "shadow rate of exchange" applied to the project cannot optimally offset the distortion in relative factor prices and can only be correct in the case of fixed coefficients Corden's statement (p 395) that the Little-Mirrlees and shadow exchange rate methods are equivalent is therefore strictly true only for the case of fixed coefficients

In their own treatment of the problem of deciding whether or not and by what technique to produce a new good, Little and Mirrlees (1974, p 178) treat production of the good with each technique as a separate project and select the one with the highest net present value in terms of world prices In principle, their solution and ours will give the same result We are able to be more explicit about the nature of the solution since we have specified a simple structure for the economy into which the project is to be introduced Unlike Corden and ourselves, however, they do not go into the question of the nature of the intervention needed to make the project viable if it is to be accepted

In yet another exposition of their approach, Little and Mirrlees (1973, p 262) state that it is equivalent to the shadow exchange rate method, differing only in the choice of the numeraire, which of course contradicts our own contention above What they apparently mean, however, is that the same assumptions must lead to the same results, independent of whether calculations are done in terms of foreign exchange or domestic currency, which is, of course, true But their statement can be misleading if it is interpreted as implying that project evaluation in a distorted open economy can be effectively carried out simply by using a single appropriate shadow exchange rate It is sometimes thought that foreign-trade taxes and subsidies only distort the foreign-exchange market, and so, in the absence of *direct* distortions in factor markets (such as union-induced wage differentials, imperfections in the capital market, etc), all that needs to be done is to get the exchange rate right As we hope to have shown conclusively, however, factor markets can be perfect, but factor prices may nevertheless be badly out of line with what they should be for a second-best optimum

II

An alternative criterion to use in evaluating a project at world prices is whether or not it could be viable in a situation of general free trade, as recommended by Bacha and Taylor (1971) This would clearly be an appealing procedure if trade liberalization were contemplated for the not too distant future In this case, the prices of outputs continue to be the world prices, but the relevant shadow prices for the factors are those that would prevail in free-trade equilibrium instead of the second-best shadow prices derived in the previous section

While the difference between these two approaches is recognized, there are some interesting questions about the relationships between them that have not been adequately answered or even posed For instance: (1) Will projects approved by the Little-Mirrlees method also be viable under free trade? (2) Under what circumstances would the Little-Mirrlees and Bacha-Taylor approaches coincide? (3) Will a project that would be viable under free trade always be accepted by the Little-Mirrlees method?

We attempt to resolve these questions in what follows The difference between the market factor prices and the second-best and free-trade shadow factor prices can be compactly represented in matrix-vector notation Let A_0 and A^* denote the technology matrices under restricted and free trade, respectively, and $\bar{p}$ and p^* the market and world product price vectors We then have $\bar{w} = A_0^{-1}\bar{p}$ for the market factor prices, $w^0 = A_0^{-1}p^*$ for the second-best shadow factor prices, and $w^* = A^{*-1}p^*$ for the free-trade shadow factor prices

The Bacha-Taylor method requires the calculation of w_1^* and w_2^* from the above relationship, and a project will be accepted or rejected depending upon whether $p_3^* \lesseqgtr w_1^*a_{13}^* + w_2^*a_{23}^*$, where a_{13}^* and a_{23}^* are the technical coefficients that minimize the cost of producing a unit of the new commodity at the w^* factor-price vector It is immediately apparent that the Little-Mirrlees and Bacha-Taylor approaches coincide if and only if A_0 and A^* are identical, which would be the case if there are fixed technical coefficients in *existing* production, both methods will select the same technique for the new good if the possibility of choice exists in its production This is a useful result if, as seems quite plausible, technical coefficients are variable ex ante and fixed ex post The two methods are equivalent in this case, and the answers to our first and third questions will both be in the affirmative In general, however, A_0 and A^* will not coincide, and then both questions remain open It will now be convenient to proceed with a diagrammatic analysis for the further resolution of these questions

The relationship between the three sets of factor prices is depicted geometrically in the Lerner diagram of figure 1 The *FF* is the factor price line corresponding to the free-trade situation, with the isoquants

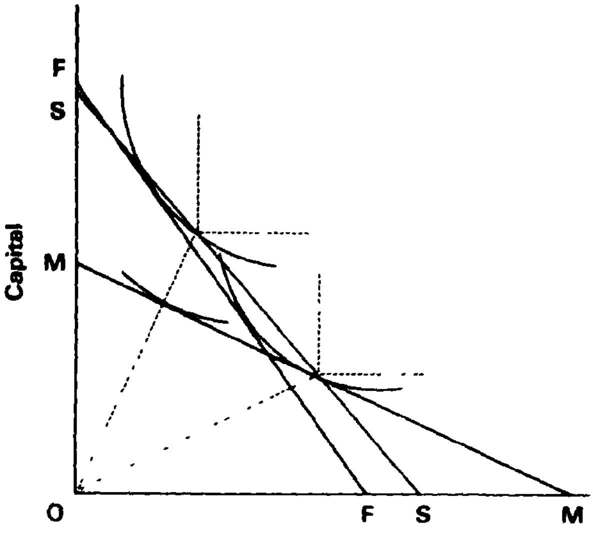


Fig 1

for the two goods representing output levels that have the same value at world prices Line *MM* corresponds to the actual market factor price line, with the isoquants tangential to it representing output levels that have the same value at internal prices with less of the capital-intensive good being worth the same amount of the other good as a result of the protection it receives The price of capital is higher and the price of labor lower in accordance with the Stolper-Samuelson theorem, and both capital-labor ratios will fall Line *SS* represents the shadow factor prices corresponding to the same capital-labor ratios as in the protection situation but with relative product prices being the same as under free trade, so that we are on the same pair of isoquants as in that situation

Both *FF* and *SS* are steeper than *MM*, indicating a lower price of capital and a higher price of labor In comparison with each other, however, it is not possible to say whether *FF* is steeper than *SS* or the other way around The outcome depends on the relative elasticities of substitution in the two sectors This is clearly brought out by considering two extreme situations First, suppose that the elasticity of substitution in the capital-intensive sector were zero so that the isoquant tangential to *FF* is L-shaped, while the isoquant for the labor-intensive good is convex In this case it is readily seen that the slope of *SS* must be in between that of *MM* and *FF* If the situation were reversed, however, with zero elasticity of substitution in the labor-intensive good, then *SS* would be steeper than *FF*, which is of course always steeper than *MM*

We now investigate the relationship between the second-best criterion of requiring a new project to increase the value of total output at world prices with protection continuing to be given to one of the existing goods and the criterion of requiring the new project to be viable in a situation of general free trade A project would be viable if under free trade the isoquant for the new good having the same value at world prices as the isoquants for the two existing goods tangential to *FF* does not lie entirely outside *FF* In other words, this isoquant must either cut *FF* or be tangential to it at some point for the project to be viable under free trade On the second-best criterion the project passes if this same isoquant cuts *SS* or is tangential to it at some point

If *FF* and *SS* intersect, then it is clear that a project may (*a*) pass both tests, (*b*) fail both tests, (*c*) pass the second-best test and fail the free-trade test, or (*d*) pass the free-trade test and fail the second-best test

If *FF* were to lie entirely inside *SS*, as is possible, the last case would be ruled out Any project that would be viable under free trade would be accepted under the Little-Mirrlees approach Another way to express this situation is to say that passing the Little-Mirrlees second-best test is necessary but not sufficient for passing the Bacha-Taylor free-trade test, while passing the latter is sufficient but not necessary for passing the former

The fact that a project accepted by the Little-Mirrlees approach might not be viable under free trade may not be sufficiently realized, and Little and Mirrlees do not explicitly draw attention to it as far as we are aware The unwary might all too easily imagine that a project which had passed a world-price test with flying colors must be able to stand on its own under free trade As our previous analysis has shown, however, this is strictly true only if there are fixed technical coefficients in all sectors of existing production

Possibility (*d*), that a project which would be viable under free trade could reduce the value of total output at world prices in the distorted situation, is certainly not intuitive and has not been noted anywhere to the best of our knowledge Passing the Bacha-Taylor test is not sufficient, as well as of course not necessary, for passing the Little-Mirrlees test, except for the situation in which *FF* lies entirely inside *SS* It is not possible for the reverse situation, *SS* entirely inside *FF*, to occur This means that it is impossible to have a situation in which passing the free-trade test is necessary but not sufficient for passing the second-best test, or in which passing the second-best test is sufficient but not necessary for passing the free-trade test

If the Bacha-Taylor criterion were adopted, the optimal subsidy-tax vector for factor prices would be ($\bar{w} - w^*$) since this would induce private producers to use the correct free-trade factor prices w^* instead of the market prices $\bar{w}$ in their choice of factor proportions for any new good. For exactly the same reason that a second-best shadow exchange rate or an output subsidy will not be the optimal form of intervention to implement project evaluation à la Little-Mirrlees, a "free-trade equilibrium" exchange rate or an output subsidy will not be optimal for the Bacha-Taylor criterion unless there are fixed coefficients of production for the new good

The model that we have used in this paper is a very simple one, but much more generality can be introduced without altering the logic of the analysis. It is readily apparent that the case of n primary factors and n tradable goods only requires us to alter the dimensions of the technology matrices and the product and factor-price vectors In principle our methods can be extended to cover all the disaggregation of the multisectoral programming models, which will be necessary for any realistic empirical application of the cost-benefit methods we have been examining In this connection it is curious to note that Little and Mirrlees express some hostility to the mathematical programming approach and give the impression that their world-price method is somehow much simpler and easier to implement In our opinion, however, a consistent and exact application of the Little-Mirrlees method would require the same amount of information and computation as the large models that they criticize Our simple linear programming methods applied to their criterion help to bring out the essential unity of these seemingly different approaches to project evaluation in distorted open economies The world-price approach undoubtedly simplifies and clarifies the general principles underlying the application of cost-benefit analysis, but it is an illusion to think that it therefore economizes on data and calculation as well It is simple and illuminating to say that the shadow price of an input is just the foreign exchange value of the decline in total production that would result if there were one unit less of that input It is no easy matter, however, to translate this principle into a meaningful number, since the sectoral composition of the change in physical production has to be estimated, and this is just what one does in a programming exercise

We conclude by summarizing the results of our analysis, which are as follows

1 Second-best shadow factor prices have been explicitly derived and compared with market prices and free-trade prices

2 The optimal way to offset the effects of distortions in the relative prices of existing goods on the production of a new good is by subsidies and taxes on the market prices of factors to make them correspond to the relevant shadow prices

3 Output subsidies or exchange-rate adjustments are not the optimal form of intervention except when there are fixed coefficients of production for the new good

4 Second-best and free-trade shadow factor prices are identical if and only if there are fixed technical coefficients in existing production

5. When technical coefficients are variable, projects that would be viable under free trade may reduce welfare if distortions continue, while projects that would improve welfare under these circumstances may not be viable if trade is liberalized This shows how sensitively cost-benefit calculations depend on assumptions made about the future course of trade policy, which is usually highly uncertain in most developing countries, thus pointing to a painful dilemma for the project evaluator

6 There is a basic unity to the world-price and programming approaches to project evaluation, the data and computational requirements for a complete analysis being the same in both cases.

References

Bacha, E , and Taylor, L "Foreign Exchange Shadow Prices A Critical Review of Current Theories " *Q J E* 85 (May 1971) 197–224

Bruno, M "The Optimal Selection of Export-promoting and Import-substituting Projects " In *Planning the External Sector Techniques, Problems and Policies* New York United Nations, 1967

Chenery, H "Comparative Advantage and Development Policy " *A E R* 51 (March 1961). 18–51

Corden, W M *Trade Policy and Economic Welfare* Cambridge Oxford Univ Press, 1974

Dasgupta, P , and Stiglitz, J "Benefit-Cost Analysis and Trade Policies " *J P E* 82, no 1 (January/February 1974) 1–33

Joshi, V "The Rationale and Relevance of the Little-Mirrlees Criterion " *Bull Oxford Inst Econ and Statis* 34 (February 1972) 3–33

Lal, D *Methods of Project Analysis A Review* World Bank Occasional Paper no 16, Washington, D C , 1974.

Little, I M D , and Mirrlees, J A *Manual of Industrial Project Analysis in Developing Countries* Vol 2 Paris Org Econ Cooperation and Development, 1969

——— "Further Reflections on the OECD Manual of Project Analysis in Developing Countries " In *Development and Planning*, edited by J Bhagwati and R Eckaus Cambridge, Mass M I T Press, 1973

——— *Project Appraisal and Planning for Developing Countries* New York Basic, 1974

National Health Insurance: Some Costs and Effects of Mandated Employee Coverage

---

Bridger M Mitchell and Charles E. Phelps

The RAND Corporation

This paper assesses some economic consequences of financing national health insurance by mandating that employers provide insurance for their employees and dependents Effects in 1975 are simulated using data from a 1970 health care survey The findings are as follows (1) Employer premiums would rise between $5 and $21 billion for plans under congressional consideration (2) Without offsetting subsidies, these premium increases will cause a transitory increase in unemployment of 0 4–1 4 percentage points (3) As employer premium payments are shifted to employees, taxable income will fall, leading to additional tax expenditures of $1 3–$5 9 billion (4) Tax expenditure on existing subsidies to health insurance is $6 4 billion

> Sir, I obey the mandate
> [*Othello* 4 1]

I. Introduction

Private health insurance is today a major industry in the United States, with annual premiums in 1975 amounting to some $29 billion Group health insurance, 96 percent of which is provided through employers and

This research was supported in part by grant HS 1029-01 from the National Center for Health Services Research, Department of Health, Education, and Welfare, and in part by the Health Insurance Study, a grant from the Department of Health, Education, and Welfare The opinions and conclusions expressed herein are solely those of the authors and should not be construed as representing the opinions or policy of any agency of the U.S government The data were generously provided to us by Ronald Andersen, Study Director of the 1970 Health Care Survey which was conducted by the Center for Health Administration Studies of the University of Chicago Helpful comments on previous drafts of this paper were given by Michael Boskin and Mark Freeland, by RAND colleagues Stanley M Besen, Craig Foch, John Koehler, and Joseph P Newhouse, and by an anonymous referee for this *Journal* We also extend our gratitude to Bryant Mori for careful computational analysis Liability for errors remaining in the paper rests solely with us A longer version of this paper has been published as RAND Corporation report R-1509-HEW (1975)

[*Journal of Political Economy*, 1976 vol 84 no 3]

© 1976 by The University of Chicago All rights reserved

labor unions, accounts for 83 percent of the total premiums, including approximately $3 billion in charges for administrative services and risk bearing [1]

While employer group health insurance has brought coverage to a large number of Americans (70 percent of all full-time employees), there is strong public interest in establishing some form of national health insurance (NHI) to make protection against health care expenses universally available One proposed approach to financing a national plan is to require by law that all employers make available health insurance policies of specified benefits and characteristics to cover all employees and their dependents This "mandated" approach would extend health insurance coverage to some 30 percent of the full-time workers who are not already covered under employer fringe benefits and would increase the benefits for most presently insured workers To be universal in coverage, mandated employer group health insurance must be combined with some other method of insuring persons who are neither dependents of nor are themselves regular full-time employees—persons who are chronically ill, disabled, unemployed, retired, or students The mandated approach to financing is the keystone of several prominent NHI proposals, including the administration's Comprehensive Health Insurance Program, and bills sponsored by the American Medical Association, the American Hospital Association, and the Chamber of Commerce (U S Department of Health, Education, and Welfare 1974*b*)

In the following section we report the increases in premium payments that employers would be required to make in order to meet the minimum standards of coverage mandated under several NHI proposals, our analysis highlights the differences that are observed across broad industrial classifications Then, using these estimates of increased premiums, we estimate in Section III the short-run reductions in employment to be anticipated because of the higher nominal costs of hiring additional workers

In Section IV we assess the extent of personal income tax subsidies of insurance that have occurred as a result of the payment of insurance premiums by employers As nontaxable fringe benefits, the compensation received in this form is exempt from personal income, corporate income, and social security taxes We estimate both the implicit tax expenditure resulting from current employer-paid premiums and the additional tax revenue that would be forgone under mandated NHI, assuming that current tax law is unchanged [2]

---

[1] These values are projections of 1970 data Methods of projection to 1975 levels and a description of the distributions of coverage and premiums are available in Mitchell and Phelps (1975)

[2] Since this study focuses on employer group insurance, we direct our assessment of the tax expenditure to that insurance only An earlier study (Mitchell and Vogel 1975)

Private health insurance and existing public programs currently pay for approximately half of all personal medical expenses, with average reimbursement rates ranging from 90 percent for hospitalization to approximately 40 percent for ambulatory care and substantially less for other services (Newhouse, Phelps, and Schwartz 1974) The proportion of costs reimbursed under an NHI plan varies widely according to the deductible and coinsurance payments required of patients and the number of services covered Some bills would result in little change, the administration bill would increase reimbursement to an estimated average of 55–60 percent, *others would provide nearly complete payment* Because administrative costs are a roughly constant percentage of benefits paid, the increases in total premiums (Sec II) provide a useful summary measure of the increased reimbursement

An NHI plan could have important consequences for social welfare by altering the distribution of risk bearing, the price, quantity, and quality of medical care, and the supply of labor (at least for workers at or near the minimum wage) These economic costs have been discussed by, for example, Arrow (1963, 1973*a*, 1973*b*), Pauly (1968, 1969), Zeckhauser (1970), Feldstein (1973), and Phelps (1973) The present paper focuses on the changes in budgetary costs for employers and taxpayers that would result from mandated financing by employers and may be considered complementary to the earlier analyses of the social costs of insurance

II. Employer Cost of Mandated Insurance

To estimate the increase in employer premiums that would occur under a mandated program, we will compare the present level of employer premium payments on behalf of each full-time employee and his family with the amounts that would result from various levels of mandated benefits The relevant parameters of mandated insurance are the total premium, T_m, needed to provide the mandated benefits and the minimum share, S_m, that would be required of employers Thus, an employer would pay $E_m = S_m T_m$ per employee electing coverage, with the employee paying the balance of $(1 - S_m) T_m$ The total premium further depends on whether coverage is for an individual or for both the employee and his family The financial parameters of a mandated plan may thus be summarized by the vector (S_m, T_m individual, T_m family) For example, for fiscal 1975 the administration bill has an average individual premium T_m = \$240 and an average family premium T_m = \$600 and requires an employer share of S_m = 0 65, which must be increased after 3 years

---

estimated the tax subsidies arising from personal income tax deductions of health insurance premiums paid directly by families, together with the direct subsidy to medical care arising from deductions of unreimbursed medical expenses to 0 75. Thus, this plan is represented by (0.65, $240, $600). Other major bills have family premiums of from $500 to $1,000 and mandate employer premium shares that range from 50 to 75 percent.[3]

Definition of Insurance Units

We have classified all individuals and families with one or more full-time workers into either *individual* or *family insurance units* in order to estimate the number of individual and family policies that will be required.

Individual insurance units consist of (1) single (unrelated) individuals working full-time, and (2) households of two full-time workers with no dependents, in which case each person is considered a separate, individual unit Under a mandated NHI plan, a husband and wife with no other dependents, both working full-time, will be able to obtain the same coverage at lower personal cost by purchasing two individual policies rather than one family policy[4]

Family insurance units consist of all other households having at least one full-time worker This category includes couples one member of which is not working full-time, families with children, and other types of families

Estimates of Increased Employer Premiums

If the insurance unit is not covered at the time NHI is mandated, the mandated employer premium payment, E_m, for that unit is equal to the mandated employer share times the total premium, that is, $E_m = S_m T_m$

However, in the case of a worker for whom an employer is already making premium payments E, the employer's contribution must increase by $E_m - E$, unless the employer payment already exceeds E_m Thus, our estimate of the increased employer payment for an insurance unit is $\Delta E = \max(E_m - E, 0)$[5]

When an insurance unit with nonworking dependents has two or more full-time workers and is now covered by group insurance policies from more than one employer, we have taken the current employer contribution, E, to be the sum of the employer contributions for all policies Our calculation, therefore, assumes that a single mandated family policy would replace the two or more policies now covering that unit About 25 percent of the insurance units classified as families fall into this category

---

[3] Under all mandated NHI plans, total premiums paid by individual firms will vary somewhat according to the age distribution and health experience of their employees All calculations in this study are based on average premium values as estimated by U S Department of Health, Education, and Welfare (1974*a*)

[4] Such a couple may at present not have two single policies (e g , one employer may not provide any coverage), and the couple may therefore be covered by a family policy from one employer

[5] Somewhat larger increases in employer premiums would result if employers maintain their current share if it now exceeds the mandated percentage, whereas a smaller increase in premiums could result if employers who now provide a very broad spectrum of services were to reduce them to those mandated Our estimates are only slightly affected by these assumptions and may be considered conservative (see Mitchell and Phelps 1975)

Data Source

Our basic data source for this study is a stratified (weighted) national probability cluster sample of the noninstitutionalized U S population in 1970.[6] The uniquely valuable feature of these data is that they include the insurance premiums paid by (or for) the sample families, obtained from verification forms sent to each insurer named by the families interviewed Thus, explicit data on total premiums, plus a breakdown of payments by employer and by family, are available, and, since the data derive from insurance company records, problems of recall error by respondents are minimized

However, for about one-third of the insured households, verified insurance premiums were unavailable, primarily because the insurers had not responded to requests for premium data We discarded those observations and then reweighted the remaining (insured) observations to correspond to the proportion of the entire sample with group insurance policies [7]

Projections to 1975 Levels

Our data source contains information on premiums of employed persons in 1970 To estimate effects of mandated NHI in 1975, we made three types of adjustments (1) To represent the "normal employment" labor force in 1975, the sample was reweighted by extrapolating employment growth rates observed from 1970 to 1973 (2) To reflect both real changes in amounts of coverage as well as nominal price increases in insurance resulting from inflation, premiums for 1970 policies were multiplied by a factor of 1 6, derived by extrapolating observed premium growth rates from 1970 to 1973 (3) To provide the basis for calculating the tax rates that would apply to employer premium payments if they were declared taxable (see Sec IV), incomes of families in 1970 were increased to 1975 levels using industry-specific growth rates observed for earnings from 1970 to 1973

---

[6] A description of the study and survey methodology is contained in Andersen et al (1972)

[7] Our estimate using this verified subsample projects an aggregate $13 13 billion in premiums in 1970, or about 3 8 percent less than the $13 65 estimate from Mueller (1972) If we included part-time workers with insurance, our aggregate estimate would rise slightly Thus, little if any bias results from the use of families with only verified premiums.

TABLE 1

INCREASE IN EMPLOYER PREMIUM PAYMENTS FOR THREE REPRESENTATIVE NHI PLANS, 1975

	Low	Intermediate	High
	Level of Mandated Employer Contributions		
Total premium			
Individual, family	$200, $500	$240, $600	$320, $1,000
Mandated employer share	50%	65%	75%
	Increase in Employer Premiums		
Increase in premium (billions)	$4.88	$8.81	$21.25
Mean increase per worker	$70	$126	$304

Results

The total increases in employer premium payments that would result from three levels of coverage—NHI plans that are representative of the range of proposals before the Congress—are summarized in table 1.[8] The low-coverage plan would have total premiums of $200 per individual and $500 per family, with an employer share of 50 percent. The high-coverage plan would have premiums of $320 for individuals and $1,000 for families and would require a 75 percent employer share. These two plans are approximately the levels of the coverage required under the bills introduced, respectively, by Representatives Fannin and Ullman. An intermediate plan, corresponding to the administration's NHI bill, has individual premiums of $240 and family premiums of $600 and requires a 65 percent employer share.[9]

For the administration bill, which mandates (0.65, $240, $600), the increase in 1975 employer premiums (family plus individual policies) is estimated at $8.8 billion. If a 75 percent employer share (required after 3 years under the administration bill) were mandated in 1975, the increase in employer premiums rises to $10.7 billion. By comparison, total 1975 health insurance premiums paid by employers are estimated to be $16 billion.

For the low-coverage bill, which mandates (0.50, $200, $500), the estimated increase in employer premiums is $4.9 billion, or about 55 percent as large as the increases under the administration bill. Most of the difference in increased premiums between the two bills is due to the reduction from 65 to 50 percent in the required employer share rather than to the difference in the mandated premium levels

---

[8] These estimates are derived from more detailed tables available in Mitchell and Phelps (1975).

[9] The bill introduced in April 1975 by Representative Fulton and supported by the American Medical Association also requires a 65 percent share; the medical services covered are similar to those of the administration bill, but lower out-of-pocket payments would be required. *Total family premiums for this bill would therefore be greater than* $600 but less than $1,000.

For the high-coverage bill, which mandates (0 75, $320, $1,000), the estimated increase in employer premiums is $21.2 billion These increases are approximately 2 4 times as large as those resulting from the administration's plan, although the total premiums required are only about two-thirds higher

Per full-time worker, the average increases in employer premiums are $70 for the low-coverage plan, $126 for the administration plan ($153 if the mandated share is 75 percent), and $304 for the high-coverage plan By comparison, average employer payments per worker in 1975 are estimated to be $231 ($330 per insured worker)

III. Short-Term Unemployment Effects of Mandated National Health Insurance

To the extent that the employer premium payments mandated by an NHI bill are larger than existing employer insurance contributions, a mandated insurance program will increase the costs to an employer of hiring each worker In this section, we demonstrate that premium increases will cause transitory but not permanent decreases in the level of employment, and we estimate the short-run unemployment generated by various NHI plans

The Long-Run Incidence of Employer Premiums

Like employer-paid payroll taxes, mandated insurance premium payments by employers are a component of a firm's total labor compensation Research on the incidence of taxation has generally concluded that payroll taxes are predominantly, if not completely, borne by labor in the long run (Pechman, Aaron, and Taussig 1968, Brittain 1971, 1972) Generalizations of the conventional theory (Boskin 1972, Feldstein 1972, 1974) establish that eventually a small proportion of the payroll tax may be borne by capital We will assume that if employers are required by law to increase their premium payments above present levels, in the future wages will change to reflect those increased costs of labor, so that the incidence of employer-paid insurance premiums will in the long run fall entirely on labor As a result, mandated insurance will have, with minor exceptions noted below, essentially no long-term effects on the aggregate level of employment [10]

---

[10] Under this assumption, the long-run substitution of tax-free insurance benefits for taxable cash wages will result in a reduction of tax revenues, an effect which we analyze in Sec IV

Methodology of Estimating Transitory Unemployment Effects

If, in the short run, wages are inflexible downward, an increase in employer premium payments would be accompanied by a period of disequilibrium during which time employers' demand for labor would be altered Since a mandated premium is a quasi-fixed cost, one that varies with the number of workers but not with the number of hours worked, employers would have some incentive to substitute overtime work for additional employees in order to minimize costs [11] Most likely, employers would accomplish this change in the employment pattern by reducing the rate of new hiring and postponing the replacement of employees who quit, substituting for their services an increase in overtime hours from the remaining employees The generally high rate of turnover in most industries (e g , 4 percent per month in manufacturing) allows this process to occur quite rapidly

Before the effective date of newly mandated insurance, we assume that unemployment is at some normal level caused by job turnover and fiscal and monetary policies Shortly before the date on which increased employer premiums are required, firms will begin to react by increasing overtime and reducing the rate of hiring, until sometime later a maximum reduction in full-time employees is reached Thereafter, the rate of unemployment gradually returns to its normal level as increases in the general price level and in worker productivity permit firms to adjust real wages, inclusive of fringe benefits, without reducing money wages [12]

The peak transitory unemployment level will be determined by the existing level of employer premium payments, the mandated premium cost, and the ability of the firm to adjust to short-run employee compensation changes To estimate the unemployment effect, we combine our estimates of the increases in employer health insurance premiums (shown by industry in table 2) with calculations of the trade-off between overtime hours and the level of fringe benefits, on the assumption that total hours of work remain constant As described in the Appendix, this trade-off is estimated on the basis of Ehrenberg's (1971) study of the cost-minimizing rate of substitution between overtime hours and the number of employees in major U S industries

Results

Table 3 presents the estimated increases in the unemployment rate created by the higher employer premiums under the three representative levels of coverage described earlier Over all industries, the unemploy-

---

[11] These cost-minimizing incentives will exist even if some firms were able to recover the increased premium costs by raising prices

[12] We neglect the much smaller changes in the permanent structure of unemployment that may result from the permanently increased proportion of total labor compensation that fringe benefits constitute

TABLE 2

Average Increase in Employer Premiums per Worker, by Industry, for Three Representative NHI Plans, 1975

	Low	Intermediate	High
	Mandated Employer Contribution		
Total premium			
Individual, family	$200, $500	$240, $600	$320, $1,000
Mandated employer share	50%	65%	75%
	Dollars per Worker (% Annual Wage Earnings)		
Industry			
Agriculture	$97 (2.1%)	$164 (3.5%)	$361 (7.7%)
Construction	81 (0.7)	139 (1.2)	326 (2.8)
Manufacturing	39 (0.4)	71 (0.6)	219 (2.0)
Transportation	55 (0.4)	110 (0.8)	296 (2.1)
Wholesale, retail trade	43 (0.5)	85 (1.0)	251 (2.8)
Finance, insurance, real estate	86 (0.8)	147 (1.4)	330 (3.1)
Service	86 (1.1)	156 (1.9)	355 (4.4)
Other	82 (0.8)	146 (1.4)	325 (3.1)
Total	$70 (0.7%)	$126 (1.2%)	$304 (2.9%)

ment rate would increase by 0.3 percentage points for a low-coverage plan, 0.6 for the intermediate plan, and as much as 1.4 for a high-coverage plan. Ehrenberg has no estimate of the substitution effect for agriculture, but because a larger increase in employment premiums is required in agriculture than in any other industry and because average wages are low, the unemployment effects could be substantial. If an average trade-off level for other industries is assumed, the unemployment increases are 0.9, 1.5, and 3.2 percentage points, respectively, for the three representative plans. Other sectors with relatively large unemployment effects are construction, wholesale and retail trade, and services; in these industries the unemployment effect would range from 0.4 to 1 percentage point for the intermediate plan and would rise to 1.1–2.1 percentage points for the high-coverage plan.[13]

A lower-bound estimate of the time that must pass before firms succeed in shifting the increased premiums to workers and unemployment returns to normal levels may be obtained if we assume that firms are able to hold nominal wages constant from the date that new premiums are mandated

[13] If one were to assume that all but unionized firms could instantly adjust nominal wages by the amount of increased premiums, the total unemployment effects would be approximately one-third those reported. If national health insurance were financed by payroll taxes rather than by mandated premiums, similar but less pronounced employment effects would result. For persons whose wage income is above the maximum taxable earnings, an increase in the payroll tax rate has the same effect as an increase in mandated premium payments, since in each case the cost of overtime is unchanged. But for workers at lower earnings levels, payroll taxes increase the costs of regular time and overtime in identical proportions, so that little if any incentive exists for employers to substitute overtime for new hires at lower wage levels.

TABLE 3

INCREASES IN SHORT-TERM UNEMPLOYMENT RATES RESULTING FROM MANDATED PREMIUMS IN THE ABSENCE OF OFFSETTING SUBSIDIES TO EMPLOYERS

	Low	Intermediate	High
	Mandated Employer Contribution		
Total premium			
Individual, family	\$200, \$500	\$240, \$600	\$320, \$1,000
Mandated employer share	50%	65%	75%
	Percentage Point Increases in Unemployment Rate		
Industry			
Agriculture*	0.9%	1.5%	3.2%
Construction	0.4	0.7	1.5
Manufacturing	0.2	0.3	0.9
Transportation, communication	0.3	0.7	1.7
Wholesale, retail trade	0.2	0.4	1.1
Finance, insurance, real estate	0.3	0.4	0.9
Service	0.6	1.0	2.1
Other	0.4	0.7	1.5
Total	0.3%	0.6%	1.4%

* We have no estimate of employment-overtime substitution possibilities in agriculture, so we assume an average value from other industries.

In that case, the real price of labor services will be eroded over time at a rate which is the sum of the rate of inflation and the rate of increase in labor productivity. For the high-coverage plan, the average increase in mandated premiums is 2.9 percent of wages, an amount that under this assumption would be recouped in 4–5 months of no nominal wage increases if inflation and productivity increase at a combined 8 percent annual rate. The service and agricultural industries would require from 6 months to a year for equilibration.

While the unemployment effect of mandated NHI is temporary, an accompanying change in the structure of demand for employees will have important longer-run consequences for some labor groups. By requiring insurance coverage of part-time workers who are employed more than a specified number of hours per week or year, some NHI proposals would alter the relative costs of hiring part-time and full-time employees. Under such bills, a shift toward increased use of full-time employees by cost-minimizing employers will result in some permanent increases in unemployment of part-time workers, including women and members of minority races, who constitute the largest fraction of the part-time labor force. (In response to this change in demand for labor, some part-time workers may eventually shift to full-time status and thus reduce the disparity between effective wage rates for part- and full-time employees.) Conversely, if mandated NHI included *only* full-time workers, then the employment of part-time labor would become a cheaper method of production for some firms, and mandated insurance would have the reverse effects on employment and eventual mode of labor force participation of those described

Finally, in the case of workers at or near the minimum wage, there may be no possibility to shift the mandated premium increase back to the employee, particularly if the magnitude of other fringe benefits is so small that they cannot be sufficiently reduced For such individuals, the increase in unemployment resulting from the mandated premiums would be permanent

Subsidies to Employers

The estimates reported above suggest that significant transitory increases in unemployment would result from mandating at least the intermediate and high levels of coverage specified by several NHI proposals To mitigate this effect, several NHI bills provide for explicit subsidies to employers whose premium increases exceed a certain amount, typically 3 or 4 percent of average wage payments

We have calculated the unemployment effect generated by the NHI plans previously discussed at varying levels of subsidies Subsidies for increased premium payments in excess of 3 percent of firms' total wage payments have only a minimal effect in reducing the increase in unemployment for the low- and intermediate-cost plans and somewhat greater effectiveness for the high-cost plan In contrast, subsidizing all increased premiums in excess of 1 or 2 percent of firms' total wage payments will substantially reduce transitory unemployment However, the annual budget costs of such a policy are as much as $16 billion for the high-coverage plan [14]

Subsidies extending over several years will be needed to avoid an increase in unemployment only if firms are unable to shift the cost of the new premiums to employees or if they are able to do so only over many years However, historical trends in inflation and productivity suggest that the transitory unemployment effect would be more short-lived and that a more rapidly phased-out subsidy would be equally effective and substantially less costly

IV. Tax Policy and Changes in Tax Revenue

New Tax Expenditures from Mandated NHI

Employer payments of health insurance premiums are excluded from corporate income, personal income, and social security taxes. Unless the definition of taxable income is changed, the increase in employer premiums resulting from a mandated NHI bill will cause a reduction in income tax revenues Thus, even though mandated insurance is an "off-budget" method of financing NHI, it will nevertheless have direct effects on federal revenues[15] In this section we estimate the revenue losses, or "tax expenditures," that would occur at various levels of mandated NHI under the assumption that definitions of taxable income remain unchanged In addition, we estimate the tax expenditures that result from the current level of employer premiums, revenue which would be available to the Treasury if employer premiums were declared taxable income

---

[14] Detailed calculations are given in Mitchell and Phelps (1975).

To carry out the tax analysis, we have reclassified the observations in the sample survey to correspond to IRS personal income tax-paying units We assume that all married couples file joint returns, that unrelated individuals are taxed separately, and that persons (other than the head of household or his or her spouse) with incomes greater than $600 are separate tax units

To estimate the tax revenue forgone, we will again assume that, in the long run, employer premiums substitute for wages in total employee compensation and that the incidence of the existing payroll taxes is also fully on labor Therefore, the relevant tax rate for each family with employer-paid health insurance is the sum of the personal income marginal tax rate and any applicable payroll taxes[16]

Results

The "low-cost" mandated NHI package results in a reduction in federal tax revenues of $1 3 billion For the administration bill, the tax expenditure is estimated to be $2 4 billion (If the 75 percent mandated share were required in 1975, the estimate would rise to $3 0 billion) For a plan with the largest mandated premium ($1,000 family coverage), the tax expenditure is estimated to be $5 9 billion[17]

These magnitudes of forgone tax revenues may be compared with the direct increase in federal spending projected for each of the illustrative plans Using the cost estimates projected for 1975 by the Department of Health, Education, and Welfare, we compare in table 4 the direct and indirect budget effects of three proposed mandated NHI plans Although

---

[15] We neglect the additional state and local income tax revenues that will also be forgone

[16] We would like to thank Benjamin Okner of the Brookings Institution for providing tax rate calculations specifically for this study Methods used to estimate actual marginal tax rates used for the study and detailed tables of tax expenditures are reported in Mitchell and Phelps (1975)

[17] In her study of NHI proposals, Davis (1975) estimates substantially smaller changes in tax expenditures than those calculated here, the differences appear to derive from her assumption of a substantially lower effective tax rate

TABLE 4

CHANGES IN THE FEDERAL BUDGET FOR ILLUSTRATIVE NHI PLANS ($ BILLIONS)

Plan	Direct Expenditure*	Tax Expenditure	Total Expenditure	Tax Expenditure as % of Direct Expenditure
Fannin	$5.4	$1.3	$6.7	25%
Administration	5.9	2.4	8.3	41
Ullman	18.3	5.9	24.2	32

* U.S. Department of Health, Education, and Welfare (1974*a*).

the details of the subsidies to low-income families and other aspects of the programs differ, for each plan the additional tax expenditure is approximately 25–40 percent of the additional direct federal spending. Thus, although mandated premiums are often thought to have no impact upon the federal budget, in fact the tax revenue that is forgone is a substantial proportion of the anticipated new public expenditures.

Tax Expenditures Resulting from Existing Employer-paid Insurance

The present employer premium payments themselves generate a substantial amount of federal tax expenditure. This indirect subsidy, initially justified as an incentive to encourage the purchase of insurance, creates a transfer payment that increases with the taxpayer's income. In table 5 we show the current distribution of employer-paid premiums by 1975 income and the average subsidy per family unit.[18] Actual employer-paid premiums rise with income, in part due to the positive partial income elasticity of demand for health insurance, as well as the correlation of multiple wage earners with higher family income. On average, families and single individuals with employer group insurance receive a tax subsidy of $122, equal to 32 percent of their employers' premium payments. In aggregate, the tax expenditure in 1975 that results from excluding employer-paid health insurance premiums from the tax base is estimated to be $5.5 billion, a figure higher even than the new tax expenditure resulting from most NHI proposals.[19]

---

[18] Each observed 1970 premium has been increased by 60 percent. Incomes are increased by industry-specific growth rates, the average of which is 39 percent.

[19] Similar calculations using 1970 premiums from the survey and 1970 marginal tax rates calculated from the IRS sample yield an estimated $2.4 billion in 1970 tax expenditures. Feldstein and Allison (1972), using substantially the same tax incidence assumption, obtain an estimate of $1.63 billion for 1969. As they note, their data base requires them to assume that within each of 32 industries all employees receive the average employer contribution, and this results in some downward bias in their estimate. Our 1970 estimate of $2.4 billion suggests that the Feldstein-Allison procedure is indeed conservative.

TABLE 5

Current Distribution of Employer-paid Health Insurance Premiums by 1975 Income and Average Subsidy per Tax Unit

Income Class (1975 Income) of Tax Unit* ($)	Families with Employer Group Insurance (%) (1)	Total Family Premiums per Insured Tax Unit† ($) (2)	Employer Premiums per Insured Tax Unit ($) (3)	Average Marginal Tax Rate, 1975 Tax Law (%) (4)	Subsidy per Insured Tax Unit ($) (5)	Aggregate Subsidy ($ Billions) (6)
0–3,000	8	336	219	22 69	48	0 02
3,000–5,000	26	329	204	28 95	60	0 09
5,000–7,000	53	356	214	31 49	68	0.21
7,000–10,000	66	441	266	31 73	84	0.61
10,000–15,000	81	641	433	31 36	137	1.86
15,000–20,000	88	619	406	30 63	124	1 10
20,000–50,000	92	629	436	34 06	151	1 58
> 50,000‡	100	720	720	50 30	362	0 05
All families	70	570	378	31 82	122	5 52

---

* Tax units are families and individuals as defined for tax-paying purposes (see Sec IV)

† "Insured tax unit" means any family or individual with employer group insurance To obtain an estimate for all tax units in this income class, multiply the value shown in this column by the percentage given in col 1

‡ One observation

Health insurance premiums paid by individuals and deducted under the personal income tax also result in forgone tax revenue, estimated by Mitchell and Vogel (1975) at $0 6 billion in 1970. Projected at the same rate of increase as employer contributions, individual premiums would receive a total tax subsidy of $0 9 billion in 1975 When this is combined with the tax expenditure on employer premiums, the total federal subsidy to existing health insurance in 1975 is estimated at $6 4 billion

The mandating of NHI for all employees and their dependents would eliminate the need for the current price incentive, in the form of an indirect tax subsidy, to encourage the purchase of health insurance Redefining employer premium payments as taxable income and eliminating the personal income tax deduction for health insurance would result in an increase in federal tax revenues of about $6 billion in addition to the amounts, as shown in table 4, that would otherwise be forgone because of the tax loss on additional mandated premiums [20]

V. Summary and Conclusions

Requiring employers to provide health insurance policies of prescribed standards will have three pronounced effects analyzed in this paper. First, employers will substantially increase their health insurance premium payments In 1975 employer premiums were $16 billion, proposals before the Congress are estimated to increase these payments between $4 9 and $21 2 billion

Second, these premium increases would, in the short run and in the absence of offsetting subsidies, decrease the number of persons employed For the proposals under consideration, the range of increase in the unemployment rate is between 0 4 and 1 4 percentage points, with significant variation by industry Services, wholesale and retail trade, and agriculture are the sectors affected most If carefully designed, temporary subsidies to employers can offset these effects to a substantial degree

Third, as the short-term unemployment effects of a mandated NHI plan diminish, a long-run reduction in tax revenues will take effect as the newly required premium payments are eventually reflected in lower taxable income for employees Our estimates of the increased tax revenues that would be forgone in 1975 if mandated NHI were enacted range from $1 3 to $5 9 billion for representative proposals These tax expenditures are between 25 and 40 percent of the new direct federal outlays for those NHI proposals Although an "off-budget" financing mechanism, mandated insurance nevertheless has significant budgetary effects

---

[20] The increase in employer premiums and the reduction in out-of-pocket payments that would result from mandated insurance would somewhat reduce the personal income tax deductions for insurance premiums and large medical expenses and somewhat offset the reduction in federal tax revenues However, the magnitude of the offset is small in relation to the revenue effects of increased employer premiums

The additional tax revenue not presently collected on existing employer group health insurance premiums and on personal income tax deduction of premiums amounted to about $6 4 billion in 1975 Combining this amount with new tax expenditures from a mandated NHI plan would result in total tax expenditures on health insurance premiums in 1975 of $8–$12 billion, or about 4–6 percent of the federal revenues from personal income and payroll taxes in 1975 Elimination of these indirect subsidies, no longer needed to encourage the purchase of insurance under a mandatory program, could provide a substantial increase in federal revenues

Appendix

Estimation of Transitory Unemployment Effects

To estimate peak transitory changes in unemployment resulting from mandated premium increases, we combine our estimates of newly required premiums for each worker with estimates by Ehrenberg (1971) of the rate at which firms can substitute between overtime hours and the number of employees Anticipating stochastic fluctuations in both the demand for their products and absenteeism of their workers, firms employ labor at overtime rates in order to avoid more costly changes in their rates of production

If firms were able to substitute, hour for hour, overtime work for regular time, with no reduction in labor productivity, the number of hours required to produce a given level of output would be constant In this case, a firm would have the substitution possibilities described by a rectangular hyperbola, $Nh = L$ (constant), shown in figure 1, where N is the number of employees, h is the hours of work per year per employee, and L is total hours However, to maintain a given level of output, it is likely that firms need to make some increase in total hours when replacing workers with additional overtime Ehrenberg assumes that this is the case and that firms maintain a constant level of labor services, so that the curve S depicting these substitution possibilities lies above the rectangular hyperbola The coefficients of Ehrenberg's equations are estimated from a cross section of 4,009 firms and measure the differences in the mean amount of overtime employed by firms whose stocks of capital and production conditions have been adjusted to permanent differences in fixed labor costs However, because not all of these adjustments can take place at once, the short-run opportunity to substitute overtime for regular employees is more limited and follows the curve labeled S' [21]

The initial labor market equilibrium of a firm with fixed (fringe-benefit) costs F and an hourly wage rate w (with overtime wage rate $1\ 5w$) is at the point X_0 At this point, N_0 workers are employed and each nonabsent employee works an average of h_0 hours per week, where h_0 is typically larger than h_s, the standard number of hours per week at regular pay Newly mandated health insurance premiums cause an increase in fixed labor costs On the assumption that these

---

[21] Firms with significant variance in either demand or absenteeism will tend to adjust overtime as rapidly in the short run as in the long run, with little difference between curves S and S'

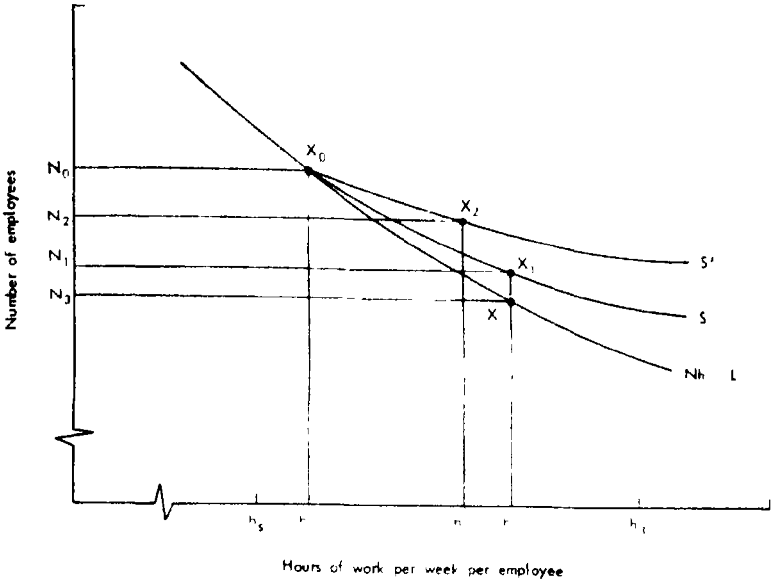


FIG 1 —The short-run labor force adjustment of the firm when fringe benefits increase

increased costs are permanent, Ehrenberg's equations predict that when sufficient time has passed for the firm to adjust other factors or production, the new equilibrium will be at X_1, with a reduction in employment to N_1 and an increase in the number of hours per worker to h_1 However, in adjusting to the short-run increase in fixed labor costs created by newly mandated premiums, cost-minimizing firms will move along the short-run substitution curve S' to a point X_2 As a result, the firm will have somewhat higher employment N_2 and shorter hours h_2 than predicted by Ehrenberg's equations In the long run, we assume that firms, by limiting increases in money wages and adjusting other fringe benefits, will reduce real wages to their initial level and return to a long-run equilibrium point on the S curve between X_0 and X_1, the precise mix of hours and employment depending on the extent to which cash wages, rather than other fringe benefits, are reduced

To calculate the transitory unemployment effects reported in the text, we have made several approximations *We first estimate the change in hours per employee* as $h_1 - h_0$, using Ehrenberg's long-run adjustment coefficients Using the value of h_1, we then calculate the change in the number of full-time employees on the assumption that total hours will remain constant, so that the estimated new level of employment is N_3 (calculated from the $Nh = L$ curve)[22] Each of these steps introduces some upward bias in our estimate of the change in the number of employed workers, so that we calculate a new employment level of N_3 rather than N_2

Offsetting these factors, however, is a bias implicit in the assumption that output will remain unchanged in the short run During the period of increased labor costs, firms are likely to reduce temporarily the level of production. In some cases, the new costs will result in higher product prices and a reduction in the quantity of output demanded. In other instances, firms will maintain prices but meet demand by drawing down inventories These adjustments will result in an inward shift of the curves in figure 1, some reduction in the demand for labor services, and therefore a smaller level of employment than shown at N_2.

---

[22] Ehrenberg makes an identical approximation to calculate the effect on employment of an increase in the overtime wage rate

To carry out these calculations, we employ Ehrenberg's industry-specific estimate of the trade-off between overtime and employment, apply our estimated increase in health insurance benefits to this formula, and then divide the predicted increase in overtime hours (OT) by average hours per year in the industry to obtain the change in unemployment (U) Thus, where Ehrenberg has estimated

$$d(OT) = \beta_j \frac{d(\text{weekly fringe costs})}{1.5 \times \text{hourly wage}},$$

we use

$$d(U) = \beta_j \frac{d(\text{weekly fringe costs})}{1.5 \times \text{hourly wage} \times \text{hours per year}},$$

where β_j is the estimated effect in industry j [23] This change in unemployment per worker is calculated for each full-time worker in the sample and aggregated to obtain the estimated percentage change in the unemployment level produced by each of the representative health insurance plans

References

Andersen, Ronald, et al *Health Service Use National Trends and Variations* Report no 1, DHEW Publication no (HSM) 73-3004 Rockville, Md Nat Center Health Services Res and Development, 1972

Arrow, Kenneth J "Uncertainty and the Welfare Economics of Medical Care" *A E R* 53, no 5 (December 1963) 941–74.

——— *Optimal Insurance and Generalized Deductibles* Report no R-1108-OEO Santa Monica, Calif RAND Corp, February 1973 (*a*)

——— *Welfare Analysis of Changes in Health Coinsurance Rates* Report no R-1281-OEO Santa Monica, Calif RAND Corp., November 1973 (*b*)

Boskin, Michael J "The Incidence of the Payroll Tax An Alternative Approach" Center for Research in Economic Growth, Research Memorandum no 136 Palo Alto, Calif Stanford Univ Press, September 1972

Brittain, John A "The Incidence of Social Security Payroll Taxes." *A E R* 61, no 1 (March 1971) 110–25

——— *The Payroll Tax for Social Security* Washington Brookings Inst., 1972

Davis, Karen *National Health Insurance Benefits, Costs and Consequences* Washington Brookings Inst, 1975

Ehrenberg, Ronald G *Fringe Benefits and Overtime Behavior.* Lexington, Mass Lexington, 1971

Feldstein, Martin S "The Incidence of the Social Security Payroll Tax Comment" *A E R* 62, no 3 (September 1972) 735–38

——— "The Welfare Loss of Excess Health Insurance." *J P E* 81, no. 2, pt. 1 (March/April 1973). 251–80.

——— "Tax Incidence in a Growing Economy with Variable Factor Supply" *Q J E* 88 (November 1974) 551–73

[23] The coefficients used are summarized in Mitchell and Phelps (1975, table 17)

Feldstein, Martin S., and Allison, Elizabeth *Tax Subsidies of Private Health Insurance: Distribution, Revenue Loss and Effects.*" Harvard Institute of Economic Research Discussion Paper no 237. Cambridge, Mass . Harvard Univ. Press, April 1972.

Mitchell, Bridger M , and Phelps, Charles E *Employer-paid Group Health Insurance and the Costs of Mandated National Coverage* Report no R-1509-NC/OEO Santa Monica, Calif RAND Corp , 1975

Mitchell, Bridger M , and Vogel, Ronald "Health and Taxes An Assessment of the Medical Deduction " *Southern Econ J* 41, no 4 (April 1975) 660–72 An earlier version is found in RAND Corporation report no R-1222-OEO, August 1973

Mueller, Marjorie Smith "Private Health Insurance in 1970 Population Coverage, Enrollment, and Financial Experience " *Soc Sec Bull* 34, no 2 (February 1972) 3–19

Newhouse, Joseph P , Phelps, Charles E , and Schwartz, William B "Policy Options and the Impact on National Health Insurance " *New England J Medicine* (June 13, 1974), pp 1345–59

Pauly, Mark V "The Economics of Moral Hazard " *A E R* 58, no 3 (June 1968) 231–37

——— "A Measure of the Welfare Cost of Health Insurance " *Health Services Res* 4 (Winter 1969) 281–91

Pechman, Joseph A , Aaron, Henry J , and Taussig, Michael K *Social Security Perspectives for Reform* Washington Brookings Inst , 1968

Phelps, Charles E *The Demand for Health Insurance A Theoretical and Empirical Investigation* Report no R-1054-OEO Santa Monica, Calif RAND Corp , July 1973

U S Department of Health, Education and Welfare "Estimated Health Expenditures under Selected National Health Insurance Bills " Mimeographed report to the Congress Washington Dept Health, Educ , and Welfare, 1974 (*a*)

——— "Provisions of Bills Introduced in the 93rd Congress as of February 1974 " In *National Health Insurance Proposals* Washington Soc Sec Admin Office Res and Statis , March 1974 (*b*)

Zeckhauser, Richard "Medical Insurance A Case Study of the Tradeoff between Risk Spreading and Appropriate Incentives " *J Econ Theory* 2, no 1 (March 1970) 10–26

The New Contractarians

Scott Gordon

Indiana University and Queen's University

One of my favorite passages in Adam Smith's *Wealth of Nations* appears near the beginning of book 1, chapter 2

> This division of labor, from which so many advantages are derived, is not originally the effect of any human wisdom, which forsees and intends that general opulence to which it gives occasion It is the necessary, though very slow and gradual, consequence of a certain propensity in human nature which has in view no such utility, the propensity to trade, barter, and exchange one thing for another
>
> [This propensity] is common to all men, and to be found in no other race of animals, which seem to know neither this nor any other species of contracts

When Adam Smith speaks of "propensity," the twentieth-century mind leaps to the interpretation that we are in the realm of behavioral or even instinct psychology, but this is incorrect, as is immediately made plain, what is unique to *Homo sapiens*, and common to all members of that species, is the making of contracts "Nobody [Smith goes on to note] ever saw a dog make a fair and deliberate exchange of one bone for another with another dog Nobody ever saw one animal by its gestures and natural cries signify to another, this is mine, that is yours, I am willing to give you this for that " Adam Smith was a true Cartesian, a philosophical mechanist up to a point The physical world consists of law-abiding particles and motions, and the organic world likewise, except for one small corner which is inhabited by that peculiar organism which thinks and values and writes books which alter the thought and values of others We are now celebrating the bicentenary of one of these great books, the *Wealth of Nations*, and while it is appropriate that we emphasize

[*Journal of Political Economy*, 1976, vol 84, no 3]
© 1976 by The University of Chicago All rights reserved

the efforts of Smith and his successors to create a science of economics in the fashion of physics and astronomy, it is also appropriate to reaffirm the central proposition that the irreducible particle of social phenomena is an entity which values and chooses and, as Smith pointed out, makes contracts with his fellows

Adam Smith's age of "enlightenment" revered John Locke as much as Isaac Newton Just as Newton's mechanics released "natural philosophy" from the hard grip of theological doctrine, Locke's theory of government similarly set "moral philosophy" free As an economist who tried to find the "laws," Smith reflected Newton, but he also reflected Locke in his contractarian view of economic phenomena Locke was not the originator of the conception, but he gave it a formulation and an expression that was propitious, aided undoubtedly by the fact that he was on the "winning side" in a society racked by civil war, regicide, military dictatorship, reaction, and adamantine religious conflict Seventeenth-century contractarian theory was an effort to find a way across this sea of troubles, and when the winds subsided and Britain resumed and accelerated her progress to constitutional democracy, that theory was firmly established as the political sextant of a rationalist age The idealist and romantic philosophical movements of the nineteenth century banished contractarian theory to the back recesses of the political cupboard, but in economic theory it remained in use, with new models of improved design being constructed to serve the needs of social navigation

We live again in a time of troubles, and the demonic winds are on the rise Leslie Stephen once remarked that the writing of political philosophy is a symptom of a recent or an impending revolution, so perhaps the exceptional amount of writing on this subject in recent years is of more than academic interest An important part of this literature is a revival of contractarian conceptions, which range all the way from banal unelaborated statements such as "we must have a new social contract" to sophisticated ethical theories which attempt to establish anew the philosophical first principles of politics The object of this paper is to discuss three major works of this latter genre John Rawls, *A Theory of Justice* (1971), Robert Nozick, *Anarchy, State, and Utopia* (1974), and James M Buchanan, *The Limits of Liberty* (1975) One of these authors, Buchanan, is a professional economist, the other two, philosophers, but a significant feature of these three books is the large role that is played in all of them by modern economic theory This attests to the fact that it is in economics that the contractarian approach to social phenomena has been sustained and developed over the past two centuries

Of the three, only Buchanan claims to be presenting a wholly "positive" theory of politics, but, like the others, he has normative conclusions and, unless one is content to commit the naturalistic fallacy, which Hume long ago exposed, one must identify the ethical principle which Buchanan employs. All three authors are alike in this respect They emply a procedural principle of ethics: the good is that which emerges from the proper procedure. This makes the sharpest philosophical contrast between our three authors and the dominant ethic of the present day which, briefly stated, is consequential that act or process is good which gives good consequences which are independently demonstrated to be so

Even if I could assume that all readers of this paper have read Rawls, Nozick, and Buchanan, it would be naive to believe that all emerge with the same interpretation of them These books are complex, and one (Nozick) would have benefited greatly from the services of a good editor Consequently, a large part of this review must be devoted to the exegesis of texts This may be useful in itself, but my main object is criticism I will indicate what seem to me to be the chief weaknesses of each of these theories as I go along, in the concluding section I will make some brief critical comments on the general approach to political philosophy which they represent

Rawls

Since its publication in 1971, *A Theory of Justice* has been the object of very extensive discussion by philosophers, political scientists, economists, and others A full examination of the issues raised by Rawls, and by his critics and defenders, is beyond the scope of this paper, I will confine myself to Rawls's basic approach to political philosophy—his effort to revive contractarianism

First, a preliminary problem of interpretation as a reader of *A Theory of Justice*, I am not completely sure of its realm of discourse—whether it is a study in moral sociology, which seeks to account for those moral judgments which are commonly held in our society, or whether it is an ethical theory, whose object is to provide a secure foundation for good moral judgments That is to say, it is not altogether clear whether Rawls is engaged in a positive or a normative investigation These are confounded, both in Rawls's own statements of his objective and in his appeal to common opinion as a support for his conclusions, but the heart of the book is assuredly normative, and I will treat it as advancing a theory of justice in this sense

Rawls starts from the position that men are self-interested, and their views of what is just are very likely to reflect considerations of personal advantage How, then, can we ascertain what is truly just? This can be done, in Rawls's view, by discovering what principles men would hold under conditions abstracted from considerations of personal advantage To determine these Rawls does not try to bring to light the will of God, the content of a benevolent human conscience, the laws of history, or the conditions for the greatest good of the greatest number He attempts instead to deduce what prudent and rational men would hold if they were in a state of perfect equality concerning their personal advantages. The principles of justice are those which would be arrived at in a hypothetical "initial condition" characterized by a "veil of ignorance" which is drawn over these advantages In such a state men are still self-interested; each wants the principles of justice to serve his interest, but all are equal in being totally ignorant concerning their relative endowments of those qualities (such as intelligence, strength, beauty, political acumen, etc) which determine degree of success in real societies

The principles of justice which would be arrived at under such conditions, according to Rawls, are not the result of compromise or consensus, this is to say, they are not the output of a political process They are the purely logical consequences of the postulated initial conditions All men are identical behind the veil of ignorance, so the principles chosen by any one would be those chosen by all, thus they are unanimous, a feature which solves (or avoids) some difficult problems in ethical theory In the same way, since the principles arrived at are the logical consequence of perfect initial conditions, they can never be improved upon and are therefore set in perpetuity, which solves (or avoids) some additional difficult problems in ethical theory

Rawls often characterizes his theory as "justice as fairness" and speaks of the principles emerging from a "fair agreement or bargain," but it is difficult to justify this characterization or his use of economic theory to amplify the thesis His initial conditions in fact reproduce the one state of affairs in which economic theory shows that no mutually advantageous bargaining is possible—where the parties are identical One could criticize Rawls along with Nozick and Buchanan as trying to reduce ethics to economics (which Frank Knight used to deplore), but, unlike the other two, Rawls reduces to the point where even economics disappears Moreover, it is more than mildly unsatisfactory, to me at least, to be presented with a package of principles that are to be regarded as the last word on the subject of justice Even physics and mathematics eschew finality, and I see no reason why we should regard ethics as more analytic than they A primary criterion of any satisfactory approach to ethical theory, especially if it is to serve as the basis of politics, is that it be "open" (in Karl Popper's sense)—that is, it ought to permit development and improvement If Rawls had demonstrated that unique definite principles of justice logically follow from the characteristics of the "initial conditions," anyone who regards "openness" as a superior value would have to reject his heuristic altogether However, though he claims such a firm analytical status for his justice principles, he does not present nor does he try to present any such rigorous derivation of them Any disciple of Rawls could describe his theory as "persuasive" but not as "compelling "

Rawls argues that any rational, self-interested individual, closeted behind the "veil of ignorance," would arrive at two basic principles of justice which are to rule as fundamental points of reference for the operation of the social order The first principle maintains the right of all members of society to an equal degree of "basic liberty", the second is the "difference principle," which maintains that inequality of actual status is permissible only to the degree that it improves the lot of the "least advantaged" members of society Rawls gives several formulations of these principles, designed to act as successive clarifications and amplifications, but I must confess that as the book goes on I become less and less confident that I understand their meaning However, the main object of this review is to examine this approach to political philosophy rather than the specifics of Rawls's theory, so I will not enter the debate on the adequacy of Rawls's two principles or his claim that the first has an absolute priority over the second The approach is deficient because it does not logically generate Rawls's principles, nor can it so generate any specific set of such principles Behind the veil of ignorance men are identical in being ignorant of their personal endowments, but are they also identical in their political and moral preferences? If not, then one can easily imagine a rational individual opting for a society in which there are a few big prizes which are randomly allocated, even at the expense of the least advantaged, or one can imagine him opting for a floor income rather than, or in addition to, Rawls's difference principle, or one can imagine him opting for a minimum degree of intergenerational or intra-life span social and/or economic mobility, and so on Rawls's method generates his specific justice principles only if we are prepared to assume that all who are closeted behind the veil of ignorance have identical views of what constitutes a just society, moreover, those views would have to be precisely Rawls's justice principles themselves! It is this identity of philosophical preferences which controls the argument, not, as Rawls asserts, the assumptions of self-interest and rationality

In short, Rawls's theory either stacks the deck by using the conclusions of the argument as its premises, or it "solves" the problem of justice only by eliminating it —by postulating a world in which there is no conflict, either of interests or of values This, I think, reflects a defect that is inherent in any theory which attempts to construct a social contract that is a collective decision What is the nature of this "decision"? If it is a vague "consensus," if it is a specific decision rule which is less than unanimity, or if it is specific, unanimous, but impermanent, then most of the problems of justice remain unsolved If, however, the collective decision is to be specific, unanimous, and permanent, it is postulated upon conditions which simply assume away the problems which are to be solved The theory eliminates what it set out to illuminate Wittgenstein once claimed to have solved the problem of justice by showing that it was (merely) a "verbal problem " Rawls does so by analyzing a condition of no conflict I cannot see that the latter is any more of a "solution" than the former, even though one may freely grant that the linguistic analysis of Wittgenstein's school and the social contract analysis of Rawls have contributed significantly to the modern discussion of political philosophy

Nozick and Buchanan cannot be accused of assuming away conflict. Their theories are, indeed, based upon conflict For them, the essential issue of political philosophy is not the derivation of principles of justice but the determination of the moral limits of state power in a world where conflict is the basic condition of life

Nozick

Of the three books that are here under review, Nozick's *Anarchy, State, and Utopia* makes the greatest demands upon the reader The basic argument is simple, but it is supported by philosophical analysis which is complex, tenacious, and profound One follows Nozick's argument with great admiration for his skill, but it is like observing the work of an archaeologist ever so skillfully peeling away the layers of protective covering from an ancient artifact, which finally appears for a moment as a thing of great beauty and then crumbles into dust from exposure to the light and air of man's ordinary environment One is compelled to distinguish between the process and its product—the one superb, the other trivial My metaphor here is intended to be more than a literary conceit, for the heart of Nozick's political theory is the proposition that the moral merit of any condition of affairs is determined by the process which led to it rather than by any qualities of that condition in itself Thus he rejects totally any "end-state principles" of political philosophy and advances instead a philosophy based on "historical principles "

The "history" which constitutes Nozick's theory is not, of course, the history of any real society (which creates great problems if one were to try to apply it in any such society), it is the hypothetical history of the classical contractarians of the seventeenth century Nozick starts his history into motion, as Hobbes and Locke did, from the "state of nature," or pure anarchy In such a condition, man is an individualist in that he does not recognize the existence of any social entity which can make legitimate claims upon his conduct But the individualism is not immaculate, each person acknowledges that other persons have the same basic rights as himself Nozick's anarchy, unlike Buchanan's, is a moral condition The mutual recognition of rights in the state of nature is not sufficient to produce a peaceful society. The boundary between mine and thine being unclear, life would be characterized by endless conflict. real or imagined injustices, disputes, retaliations, feuds—in short, a war of each against all The unhappy residents of such a habitude do not, as

Locke envisaged, agree collectively to establish a social order Each tries to protect himself as best he can, but the virtues of division of labor apply to protective activities as well as to others, so "firms" will be created, offering protective services to those who become clients Disputes among the clients of a particular firm are not eliminated, however, so firms will find it to be a condition of commercial success that they develop effective and acceptable procedures for the settlement of contending claims

Up to this point no truly significant advance over the state of nature has been made, there is peace among the clients of a particular protective agency, but there may be many agencies still in unremitting warfare against one another in any geographical area Is there any factor which will lead to the development of a monopolistic protective agency? Are there any reasons why such an agency, if it were to emerge, should supply its services to all residents, leaving no one outside its orbit? Nozick answers "yes" to both of these crucial questions Thus, a monopolistic and ubiquitous protective agency will come into existence—that is, a "state "

Nozick's argument that a monopoly of protective services will emerge from voluntary contract is, in essence, based on the proposition that there are great economies of scale in such an industry A protective firm will not defend its clients against all actions of outsiders since some such actions may reflect just claims, but, even though the firm is morally constrained in this way, it must determine what is a just claim by an outsider in the same way that it must adjudicate the contending claims of its own clients Everyone has a moral right to perform this judicial function, but there are obvious practical necessities which transfer this function from the individual to the protective agency If one assumes that such firms are not identically successful in obtaining clients, the "dominant agency" has a strong market advantage merely by being dominant "The dominant agency can offer its customers a guarantee that no other agency can match 'Only those procedures *we* deem appropriate will be used on our customers' " (Nozick 1974, p 109) Thus, an agency having a monopoly of protective services emerges from the mechanism of private and voluntary contracting This Nozick calls the "ultraminimal state "

What is now the status of any persons who wish to remain independent and decline to become clients of the protective agency? It would be morally improper for the agency to disregard their rights, so it is obligated to treat them the same as it does its contracting clients The agency must either supply its services to independents without charge, which would render the whole arrangement unstable on account of the "free-rider" problem, or it could alter its fiscal arrangements to create, in effect, a system of coercive taxation With the adoption of taxation and the supplying of protective services to all, there comes into existence what Nozick calls the "minimal state " Having shown that such a state will emerge by legitimate steps from an anarchy in which all persons are free of obligation to any social institution, the remainder of Nozick's analysis is an effort to prove that any development of coercive power beyond the minimal state is illegitimate, and, consequently, such state functions which occur in the real world are illicit and should be dismantled.

The minimal state offers only one public good—protection, and its concomitant judicial services—and engages in redistribution of income only to the extent that is necessary to carry on such functions Nozick's thesis is that one cannot describe a process whereby any other functions may be legitimately added to the state, regardless of what might be their service to general well-being The heart of this argument is that such additional functions would involve redistribution, and this renders them illegitimate If a "public good" can be financed by voluntary subscription, there is no reason why private commerical arrangements cannot provide it What about a public good for which no mechanism of exclusion is possible—for example, street cleaning? It is a good which may be desired by all members of the community, but it is crucially exposed to the free-rider problem Each person's interest lies in disguising his desire *for it and refusing any voluntary contribution, so it cannot be provided* by private enterprise arrangements, and, if we follow Nozick, it would not be provided at all This conclusion is difficult to accept One feels, albeit intuitively, that a philosophy which condemns one on grounds of justice to live with filthy streets even though everyone would prefer some quantity of street cleaning to other goods of equal cost is somehow wrong One might say that the free-rider problem arises out of an immoral act, that is, disguising one's true preferences, but if we were to conclude that the state (or another street-cleaning agency) was thereby entitled to coerce payment, the door would be opened wide to the growth of state functions far beyond the "minimal" Quite apart from the free-rider problem, difficulties arise because the nature of street cleaning is such that the same amount must be provided to all community members If they value it differently, a uniform price would be redistributive, since no person can vary the quantity of it so as to keep the marginal rate of substitution between it and other goods equal to its relative price We should have to require that each person contribute in proportion to the marginal rate of substitution of street cleaning to him, as assessed by himself Nozick's just society would either have dirty streets to a degree desired by no one, or it would have to be inhabited by men whose moral sense included much more than a recognition of the right of others to noninterference Like all utopianists, Nozick constructs an ideal society which can work only if it is inhabited by ideal persons

Nozick does not discuss the public goods problem, which is a conspicuous weakness of the book, but it is clear that the central ethical issue it raises for him is that of redistribution His argument that any state functions beyond the "minimal" are illegitimate is based upon a condemnation of redistribution (other than what might be required to rectify injustice—the door swings wide open again), so it is to this issue that we must direct our attention

The ethical quality of an income distribution, in Nozick's view, depends entirely on how that distribution came to be what it is If it arises from a morally legitimate process, it is morally just regardless of its degree of inequality or any other pattern characteristics it may possess On this ground he rejects Rawls's "difference principle" as an effort to use an "end-state" or pattern criterion as a moral rule of distributive justice He insists here, as in his theory of the minimal state, that the only valid criterion of what is morally good is a "historical" criterion If we start from a state of justice and proceed from it, by legitimate steps only, then whatever emerges is also just If, in the initial state, people are morally entitled to their holdings, then that entitlement is preserved intact when one moves from this initial state by legitimate steps The moral proposition that one has the right to do what one wills with one's own, as long as one does not interfere with the similar rights of others, is the ethical fundament of Nozick's political philosophy For an individualist who takes the view (as I do) that social institutions are not primary entities and cannot be invested with rights, duties, etc (except metaphorically) this is an appealing proposition,[1] but it is empty without a theory of what legitimately constitutes "one's own " To fill this need, a theory of "entitlement," that is, a theory of property rights, is required At this point Nozick's elegant structure crumbles

So long as we confine ourselves to the issue of the transfer of property rights by free exchange, we are in no great difficulty Nozick recognizes the contribution that economics has made to this aspect of ethical theory, but he also recognizes that what results from exchange is also dependent upon the initial distribution of holdings The theory of entitlement must provide us with a justification of some such initial distribution (which justification, as Nozick recognizes, would also act as a proper principle of "rectification" in any case where this initial distribution is unjust) What, then, constitutes entitlement in the acquisition of property prior to voluntary exchange? Nozick discusses this at length, but he does not offer an answer to this critical question I do not mean to say that he offers an inadequate theory of entitlement, he does not offer a theory at all

---

[1] It is not, however, satisfactory in itself, because it recognizes only an obligation to refrain from doing harm to others That there may be positive obligations as well is, however, difficult to deny The Samaritan was called "good" because he assisted another in need and did not neutrally "pass by on the other side " The biblical story serves to make the point that the content of the moral good is not exhausted by the absence of moral evil The recognition of positive moral obligations is, admittedly, fraught with great danger, through that small opening much tyranny may be slipped, but this requires that we exercise vigilance and balanced judgment at this point in our social (and, indeed, personal) policies, rather than an effort, which in any case is doomed to failure, to seal it off altogether

One can find in *Anarchy, State, and Utopia* a series of suggestions as to what would constitute entitlement (some of which simply beg the issue, e g , that one is entitled to the marginal product of one's labor and property!), but, just as in reading Rawls one turns over successive pages in expectation of the proof that his justice principles follow logically from his initial conditions, so one turns over Nozick's pages in expectation of a theory of entitlement In both cases the expectation remains unfulfilled Nozick's political philosophy is structured to rest upon a single pillar, which is missing One hesitates to come to this conclusion about a book of such profound intellectual power which claims so much, but Nozick himself removes misgivings on this point

> The general outlines of the theory of justice in holdings are that the holdings of a person are just if he is entitled to them by the principles of justice in acquisition and transfer, or by the principle of rectification of injustice (as specified by the first two principles) If each person's holdings are just, then the total set (distribution) of holdings is just To turn these general outlines into a specific theory we would have to specify the details of each of the three principles of justice in holdings the principle of acquisition of holdings, the principle of transfer of holdings, and the principle of rectification of violations of the first two principles *I shall not attempt that task here* [Nozick 1974, p 153, italics mine]

The final word of this passage seems to promise a further work which will provide the missing element of Nozick's political philosophy One would think that such an element, so fundamental to the argument, should be forged before committing oneself to the rest of the structure But that is perhaps a carping criticism, more important is whether it is possible to construct a satisfactory theory of property rights based on the concept of entitlement I do not think so, not because the idea of "entitlement" is utterly worthless as an ethical guide, but because it cannot serve, as Nozick would have it do, as the sole ethical guide Even if Nozick should be able in a future work to provide a theory of entitlement which advances on Locke's, it would only be one of the pillars of justice and must share the burden of carrying the house of polity with other great ethical principles, such as "utility," "freedom," "equality," and Rawls's "fairness " The architect of a good polity makes a fatal error if he seeks to build in soaring cantilever on a single shaft If we are prepared to learn from history (as well as from philosophy), it seems evident that the best political houses are rambling structures, built modestly low, on numerous footings, with not much reference to any master plan, they are not elegant at all (unless one's tastes are wildly eclectic) but variform, even motley, and, above all, easy to alter and repair

Buchanan

The Limits of Liberty is concerned mainly with two topics One is an attempt to construct a new contractarian theory of the state, and the other deals with the issue of whether there are forces which tend to make a state grow beyond its legitimate limits The latter is a matter of great practical importance and is of no small significance from the standpoint of political philosophy, but I will confine this discussion to Buchanan's effort to describe the state as based upon a "constitutional contract "

Like Nozick, Buchanan begins with an analysis of the condition of anarchy He does not arrive at an absolute doctrine of the minimal state as Nozick does, and indeed he rejects as "moral arrogance" any effort to describe the characteristics of a good society, but his basic approach to the ethical issues of political philosophy is the same as Nozick's He makes the same distinction between what Nozick calls "end-state" and "historical" principles, rejecting the first and embracing the second "My natural proclivity as an economist is to place ultimate value on process or procedure, and by implication to define as 'good' that which emerges from agreement among free men, independently of intrinsic evaluation of the outcome itself" (p 167) The process or procedure on which "ultimate value" is placed is that which starts from a state of unconstrained freedom In his opening discussion Buchanan asserts that he does not adopt freedom as a "metaphysically superior value" (as Nozick does) but recognizes it "much more simply as a necessary consequence of an individualist-democratic methodology" (p 2) This may help to explain why Buchanan describes his approach to political philosophy as the "natural proclivity [of] an economist," but without, I hope, shedding my own credentials, I have to confess that I find this confusing The issue that opens at this point is the relationship between scientific methodology and moral values I cannot enter this great debate here, but it seems to me that Buchanan is striving to commit what G E Moore called the "naturalistic fallacy" deriving moral principles without the aid of any moral premise As one of the pioneers of the modern positive analysis of the political processes of collective decision making, Buchanan attempts in this book to carry over the same approaches to fundamental political philosophy The attempt cannot succeed In a positive analysis of political processes one's assumptions may be confined to the behavioral, and the test of the theory is ultimately empirical, that is, its explanatory and predictive power, but, try as one will, that troublesome word "ought" cannot be excised from political philosophy and no degree of sophisticated "is" can take its place

Buchanan's anarchy is more individualistic than Nozick's in that the actions of men are not considered to be bounded by any recognition of the rights of others. In such a state, men will be engaged in predatory action upon one another without moral constraint. Robbery is a vehicle of acquisition, no different from production from the individual point of view, and will be carried to the point where its marginal cost equals its marginal yield. In such a world, protection against predation is also economic and will similarly be engaged in by each individual up to the point where its marginal cost is equal to its marginal yield. It is a Hobbesian anarchy wherein "he shall take who has the power and he shall keep who can." Utilizing the analysis of the late Winston Bush, to whom the book is dedicated, Buchanan arrives at the conclusion that this state of anarchy will generate an equilibrium, defined by the trim of the appropriate marginals, and the distribution of property which emerges is called the "natural distribution." Since the distribution of those personal endowments which are useful in predation and protection activity need not correspond to the distribution of those useful in production activity, there is no reason why the "natural distribution" of property should not depart markedly from that which would prevail if each person kept what he produced. The error of much political philosophy in Buchanan's view lies in starting from the distribution that would prevail with no predation or protection instead of the "natural distribution." The latter is Buchanan's archimedean point upon which he rests his conception of a "constitutional contract" establishing a political state.

In the anarchistic equilibrium considerable quantities of resources are being devoted to predation and protection. There is, therefore, opportunity to move to a higher Pareto optimum. If the natural distribution could be set by binding agreement, a large part of the predation and protection costs could be saved and devoted to production activity. (Presumably there would also be gains due to the reduction of uncertainty, but Buchanan does not make this point.) This argument, which arises purely out of self-interest without any input of moral consideration, is the "constitutional contract." Once it is established, with the appropriate instruments of enforcement, people can go about the business of private production and exchange in an environment of law and order.

Before going any further, let me say that I find the Bush-Buchanan scenario of anarchistic equilibrium unconvincing. Since there are such large gains to be made from reducing predation-protection activity, would it not be profitable for a small coalition of persons, endowed with exceptional organizational skills, to make a once-and-for-all investment in predation with the object of permanently enslaving all others? Once done, a "state" is established, law and order reigns, and one moves to a higher optimum—for the rulers at least. As a description of a man's actual history this seems more accurate than Buchanan's constitutional contract. Buchanan recognizes this possibility, but he muddies the central issue by taking the view that, under certain conditions, slavery for some would be their best option and would therefore be a rational voluntary contract on their part! A notable feature of Buchanan's and Nozick's theories is that both are prepared to recognize the legitimacy of slavery when "voluntarily" entered into Thus one finds the defenders of freedom defending the most complete form of coercion This remarkable result would delight a Hegelian, but even if one does not accept the proposition that ideas "contain" their opposites, there is ample evidence that any idea, when extended and extended, tends to become its opposite Realism becomes idealism, rationalism becomes existentialism, the political "left" finds itself on top of the political "right," democracy becomes tyranny, and freedom becomes slavery Long ago, Plato used the paradox of freedom (Cannot free men choose to be governed by a tyrant?) to attack the Athenian democracy, but liberty may suffer as much from its friends as its enemies when political philosophy is driven by "inexorable logic " Political space is curved, and political principles, when extended, turn round upon themselves Even good principles must be tempered by moderation, lest one find oneself going round the circle and defending what one abhors

Buchanan recognizes that, even if one were prepared to view the constitutional contract as a real historical event, it is not clear that it will continue to be binding on future generations Each person is born into a world already made, with its political institutions already in place Why should he accept the established order? Again, Buchanan's reasoning eschews any moral considerations of duty or obligation Each person will accept the established order if and only if he (believes that he) cannot do better by "renegotiation " Everyone has the option to make a "detour through anarchy" in which a new natural distribution of property would emerge, and from this new equilibrium, a new constitutional contract One will opt for the "detour" if one's expectation of gain is sufficient to warrant the costs If a large number of people expect to receive a net benefit from renegotiation, those who are currently favored in the distribution of wealth and power may respond by repression, but it is likely to prove more practical and prudent for them to permit a redistribution Thus a way is open for changes in the constitutional contract

In the light of this argument it is difficult to see why Buchanan views the disorders of the 1960s as negatively as he does It is clear from the *Limits of Liberty* that he regards these events as wholly bad, representing a serious decline in respect for legitimate practices and a serious dissipation of that greatest of all "public capital goods," the law But, if one follows his thesis, one could easily interpret these disorders as evidence that renegotiation expectations had departed markedly from the status-quo distribution, that numerous people were opting for a detour through anarchy—a perfectly legitimate demand for a revision of the constitutional contract Similarly, one could interpret the accommodating responses of those favored by the status quo as the rational actions of prudent men, instead of deploring them as Buchanan does

Locke's political theory was a great improvement on Hobbes's because of his recognition that a constitutional order cannot be set for all time, even if there ever was such a thing as a primordial "state of nature " It must be subject to revision, accommodating to changes in material conditions and to changes in ideas Buchanan does well to follow Locke rather than Hobbes in his view of the nature of the constitutional contract, but his theory gives us no guide on the vital issues of when and how and what to change in it Economic theory as a positive analysis of voluntary exchange is a powerful scientific instrument, and we live at a captivating time in its history, when its techniques and approaches are being extended to deal with a wider range of social issues than ever before Buchanan is one of the great pioneers of this extension, but there are limits even to what economic theory can do

General Comments

In the foregoing sections I have made some comments on what seem to me to be the chief weaknesses of each of the three books under review There remains the broader issue of the general approach to political philosophy which they exemplify A full analysis of the weaknesses of contractarian theory would be a large and complex undertaking, and it would, at bottom, have to amount to the presentation of an alternative, more satisfactory, approach to political philosophy That cannot be done here even if I were equal to the task, which I doubt But, as Nozick remarks in defense of his own incompleteness, "There is room for words on subjects other than last words" (p xii), so I will end this essay with some very brief remarks on the contractarian approach to political philosophy and the procedural theory of justice which is associated with it

Adam Smith was perceptive in pointing out that the distinctive characteristic of man as social animal is that he makes contracts, but the most important feature of that word is the final letter, which makes it plural There is a world of difference between the conception of society as consisting of contracts and the conception of it as based upon *a* contract The latter could be justified as a heuristic fiction if it led one to a more acute perception of the nature of society, but it does not Contracting (and recontracting) are not the discrete and comprehensive acts envisaged by Hobbes and Locke or by Rawls and Buchanan This activity is continuous and unending, and its substance consists of small things such as the agreement to transfer the ownership of a shirt or the use of a day's labor, not big things like the establishment of a political state or the principles of justice. Without a doubt, the social and economic order emerges from such contracts just as, say, the price vectors of goods and services emerge from the many individual acts of buying and selling One can speak of the participants in a competitive market "agreeing" upon the equilibrium price, but that is a metaphorical locution—completely false if taken to be descriptive, and misleading if one attempts to base the philosophy of politics upon it

There may be some merit in Rawls's use of the contract conception as a device for throwing light on the problem of justice construed as "fairness " If one is to have a fair game it seems desirable that the rules be determined before one begins, not changed ad hoc (and certainly not unilaterally) while it is in process Similarly, if social life is to be fair it seems desirable that the rules which provide constraints upon action should be general and disinterested, "objective" in the sense of not being designed simply to legitimize the personal advantages of particular participants So far so good, but the conception of contract does not really help much in determining "fairness " Social life is a continuous game with new participants constantly entering and old ones leaving the field There is no alternative to changing the rules in the middle of the game, in fact, one of the objectives of the game is to change the rules as we learn more about how the game works and develop our sense of what we want it to achieve We might even go so far as to impose "handicaps" on some of the players (surely a nonobjective thing to do) in order to make it a better game

Being a continuous process, social life at any point in time contains its history within it, including the effects of the "unfair" or immoral acts that have taken place in the past If one adopts Nozick's view that a current state is just if it reflects "entitlement," then one must also accept Nozick's principle of "rectification " To make our society just we would have to alter all those features of it that reflect past injustices and immoral acts To perform such rectification is a practical impossibility, and, moreover, it would only mean the piling of new injustices upon the old To make a society more just means to alter its future, not to reconstitute its past, to do this we have to consider what we want it to become, not what it was in a never-never land of the state of nature

There is no escape from the necessity of using some "end-state" criteria of the social good, even though there can be no final or perfect end state to guide us Buchanan strives to avoid this by conceiving his "constitutional contract" as being subject to "renegotiation," but it is not a renegotiation on the basis of moral principles The favored members of society will renegotiate only when they are threatened, and implicitly he argues (if indeed his book is about political philosophy at all) that they ought to renegotiate only under such threat. This is guerrilla politics and hostage politics given a moral gloss. Man is, fortunately, more civilized than that, at least in some areas of the world.

The contract theory of the state distorts more than it clarifies because it mistakes the essential problem of politics, which is to deal with the continuous, the complex, and the insoluble. The contractarian is a great simplifier who searches for discrete and final answers He finds politics vexatious and wishes to do away with it, like Dostoyevsky's Grand Inquisitor who sought to protect men from the terrible burden of moral choice But a civilized polity is not built upon a contract There never was a contract, there is no contract, and there cannot be a contract at the base of social life Its only constant is change As the Hindu says, one cannot bathe twice in the same river, one can do so only in a stagnant pool

The procedural approach to the problem of justice has many practical attractions, and large parts of our social life, including the judicial system, are in fact based upon it to a considerable degree But as a general philosophy of justice it has severe weaknesses, primarily because it is difficult (perhaps impossible) to formulate a procedural theory of justice that does not beg the main issue If one were to say, "Those consequences are just which result from a procedure," it would be evident there is something essential missing The final noun in this statement requires a qualifier No one would try to argue that any procedure will serve, for that would empty the statement of discriminating capacity One could adopt a formal or legalistic qualifier, saying for example, "Those consequences are just which result from an *established* procedure " Charles I argued at his trial that Parliament had no right to judge his conduct because to so judge a monarch was not established procedure, the Nuremberg defendants argued that the acts for which they were charged had all been in accordance with the established law of the Third Reich Obviously something is amiss here too The qualifier must have moral content The statement must read something like, "Those consequences are just which result from a *fair* procedure," or in place of "fair" we might have "good" or some other term with moral content But I cannot see that we are much advanced by couching the theory of justice in procedural terms, since we must then unpack (as the philosophers say) the concepts of "fair" or "good", this is no easier to do than to unpack the concept of "justice" itself One might as well say, "Those consequences are just which result from a *just* procedure," and in exploring that I would rather examine what makes a consequence just rather than what makes a procedure just In short, I would hold that justice is a matter, to use Nozick's terms, of applying "end-state" rather than "historical" criteria

The procedural theory of justice is part of a contractarian approach to political philosophy because it focuses the issue upon prior agreement (In Rawls's initial condition there is agreement to establish principles of justice which then control subsequent policy, in Nozick's anarchy an agreement emerges establishing a "minimal state" which then controls subsequent policy, in Buchanan's anarchy an agreement emerges to fix the "natural distribution" in a "constitutional contract" which then controls subsequent policy) There is an element of merit in the contractarian or procedural approach to justice and politics since there are clear dangers in deciding issues arbitrarily and ad hoc Charles I had a point I reject the procedural approach to political philosophy, but I would accept a philosophy which views politics as a procedure Let me try to explain this rather cryptic statement

The great illusion which pervades the history of political philosophy is the belief that one can find the first principles of social justice and, having done so, establish finally upon them the structure of a just polity The illusion lies in seeking solutions to insoluble problems The principle of justice cannot be found, for there is more than one basic principle that appeals to moral sense and for which good argument can be made (Gordon 1963) We live in a morally messy world But it is the one we are stuck with (The idea that we might make contact with the inhabitants of another planet whose moral sense is more "highly developed" than ours is merely a form of the great illusion) Social decisions must be made despite the fact that the basic social problems are inherently insoluble The simplest process is to let Charles I (or his equivalent) make them, but we all know what is wrong with that In a civilized society, where all have sovereign qualities and none are "subjects," we have to have procedures for making social decisions which are open to the play of multiform interests and preferences and multiform senses of what is just They have to be tentative and experimental, capable of being altered as we learn from our mistakes Someone once described politics as the art of the possible, it would be more correct to call it the art of *im*possible, the art of making social decisions when there are no decision rules The task of politics is that of making "second best" decisions, and it is exposed to the same grave difficulty that Lipsey and Lancaster found in welfare economics But I have no faith in any effort to meet that problem by creating perfection at a stroke The process of making social decisions can be immensely improved by philosophical and social analysis, theoretical and empirical, but it can never be made perfect or reduced to formula In the realm of politics one must revise Nozick's aphorism there is room for all words *except* last words I have no doubt that Rawls, Nozick, and Buchanan would all agree to that, for their books are in the highest tradition of open and continuing debate

References

Buchanan, James M *The Limits of Liberty: Between Anarchy and Leviathan*. Chicago · Univ Chicago Press, 1975

Gordon, H. Scott "Ideas of Economic Justice." *Daedalus* 92 (Summer 1963) 433–46

Nozick, Robert *Anarchy, State, and Utopia* New York Basic, 1974.

Rawls, John *A Theory of Justice* Cambridge, Mass · Belknap, 1971

Communications

The Effect of Unemployment Compensation on a Seasonal Industry: Agriculture

Barry R. Chiswick

Council of Economic Advisers

I. Introduction

Seasonal industries have always posed a special problem for the unemployment compensation program (O'Connor 1962; American Enterprise Institute 1974).[1] Since some off-season unemployment is anticipated on a regular basis, the involuntary nature of unemployment during the off-season is questionable. In addition, there is particular concern for the effects of unemployment compensation on the work effort (employment) of those in seasonal industries. Unemployment compensation benefits based on wage rates during the on-season may well approach or exceed the worker's productivity or wage during the off-season in that job or elsewhere, thereby discouraging off-season work. This incentive would be particularly strong for workers in low-wage industries such as agriculture since, on the average, unemployment benefits appear to replace a greater proportion of their after-tax earnings. It has been estimated, for example, that unemployment benefits replace about 70 percent of lost after-tax earnings for male family heads below 150 percent of the poverty

---

I appreciate the comments received from Carmel Ullman Chiswick, Stanley Horowitz, Dave O'Neill, and G. Edward Schuh as well as Robert Schanzmeyer's research assistance. The views expressed in this paper are solely my own and are not to be attributed to the CEA. The time series ends at September 1975, the latest monthly data available when the revision was submitted to this *Journal*.

[1] For a detailed presentation of the current features of the state unemployment compensation programs, see U.S. Department of Labor (1972) and the semiannual comparison revisions. For a summary, see U.S. Department of Labor (1975). All references in this paper to the current state programs come from these sources.

[*Journal of Political Economy*, 1976, vol. 84, no. 3]
© 1976 by The University of Chicago. All rights reserved.

line, but the after-tax replacement ratio for those above this level may be only about 40 percent (Council of Economic Advisers 1975, pp. 121–23) [2] Yet there has been no systematic empirical analysis of the effect of unemployment compensation on unemployment and employment in seasonal industries in general or agriculture in particular

This paper explores the effect on the industry unemployment rate and level of employment of the introduction in January 1975 of a nonexperience-rated unemployment compensation program that covers agricultural wage and salary employment [3] The program is described (Section II) and the hypothesized effects are developed (Section III) Regression equations are estimated (Section IV) to permit the prediction of what would have been the monthly unemployment rate and level of employment of agricultural wage and salary workers in the absence of the introduction of unemployment compensation The predicted and observed values are compared (Section V) to test the hypotheses A concluding section then follows

II. Special Unemployment Assistance

The agricultural sector was excluded from coverage in the original federal unemployment compensation legislation and subsequent amendments, and prior to January 1975 only Hawaii and Minnesota covered some agricultural wage and salary workers under their state programs The Emergency Jobs and Unemployment Assistance Act, signed into law on December 31, 1974, created a temporary program called Special Unemployment Assistance (SUA) [4] The program was to provide benefits for unemployed wage and salary workers who are ineligible under the regular state or federal unemployment compensation programs because part or all of their employment was in an industry not covered by any of these programs, but who are otherwise eligible for benefits

The eligibility provisions and weekly benefits for SUA are the same as those that apply for the regular program in the applicant's state [5] Thus, in principle, the worker must have accumulated a minimum amount of work experience; the precipitating cause of the unemployment must be involuntary, or else the claimant must show "good cause" for leaving the employer, and the claimant must be available for, be actively searching for, and accept employment "comparable" to the previous job Although SUA is administered by the states, it is fully federally funded and there is no experience rating of employers

---

[2] Unemployment benefits generally replace about one-half of pretax earnings up to a ceiling, which varied among the states from about $66 to $117 per week in 1974 The benefits are exempt from federal payroll and income taxation and generally also from state income taxation

[3] The Bureau of Labor Statistics reports data on "farm workers" as well as "agricultural wage and salary workers " The former, but not the latter, includes self-employed farmers and unpaid family workers This paper focuses exclusively on hired farm workers since the self-employed and unpaid family workers are not eligible for unemployment compensation

[4] In the June 1975 amendments, the deadline for filing claims was extended from December 1975 to March 1977, and the maximum duration of benefits was raised from 26 to 39 weeks for 1 5 weeks of benefits for each week of eligibility under the basic 26-week program

[5] The one exception is that the work-experience requirement is somewhat weaker under SUA than under many of the regular state programs

III. Hypothesized Effects of SUA

A nonexperience-rated unemployment compensation program for farm wage and salary workers which is funded out of general federal revenues is an unambiguous subsidy of periods of unemployment for hired farm workers For the same farm wage rate, more workers will supply themselves to the agricultural sector since the off-season weekly "wage" is increased The increase in labor supply in the on-season reduces agricultural wage rates and thereby increases farm employment and output during the on-season

During the off-season, however, farm employment and production decrease as the difference between the value of the marginal product of hired farm workers and the value of their next best nonfarm activity decreases It decreases because a new activity which expands the opportunity set—accepting unemployment compensation—has appeared Indeed, during the off-season the weekly unemployment benefit may approximate or exceed the weekly wage, net of taxes and work-related expenses

The difference between the on-season and off-season shifts in farm employment arises from the optimal timing of work in the year Since, by definition, marginal products are lower and employment opportunities are fewer during the off-season, and since some employment is needed to qualify for future benefits, it is efficient for both the farm owners and the hired farm workers if the latter take the unemployment benefits during the off-season [6]

Thus, the seasonality of farm employment of wage and salary workers increases Employment during the off-season decreases, the length of the off-season is widened, and employment increases in the on-season The effect on the average monthly level of employment over the year is ambiguous

The seasonality in unemployment also increases Those experienced workers previously unemployed during the off-season have no reason to alter their labor-force status However, farm workers who would otherwise be employed during the off-season either full time or part time on low-wage (marginal product) farm jobs, low-wage temporary nonfarm jobs, or jobs far from their residence have a reduced incentive to accept such interim employment The greater annual farm earnings also reduce the incentive to accept long-term nonfarm employment during one's off-season unemployment In addition, workers who would otherwise drop out of the labor force because of an inability to find employment (discouraged workers) or because of family responsibilities, retirement, etc will have an incentive to report themselves as unemployed and collect benefits, particularly during periods of seasonal unemployment All of these incentives are especially strong for the agricultural sector, since the extent to which unemployment benefits replace lost wages is greater for low-wage than for high-wage workers

---

[6] For a mathematical demonstration, see Feldstein (1976)

The incentives for increased unemployment during the on-season exist but are much weaker than for the off-season Benefits for the next off-season are contingent on working during the intervening on-season Hence the marginal revenue to the worker from a week of work in the on-season includes the value of the credit for greater weekly benefits and a longer duration of benefits in the next off-season This encourages employment in the on-season However, workers assured of employment elsewhere in the next off-season, or anticipating death, retirement, or otherwise withdrawing from the labor force during the on-season, regardless of the now stronger work incentives, may attempt to collect unemployment benefits during the on-season It should be noted, however, that it is more difficult to defend the unavailability of comparable work in the farm sector in the on-season The direction of the net effect of the introduction of SUA for on-season unemployment is ambiguous, and its magnitude is likely to be small

The prediction for unemployment can be expressed in more familiar terminology—the unemployment rate During the off-season the effect of the unemployment compensation program is to decrease employment, increase unemployment, and increase the agricultural wage and salary worker labor force These result in an unambiguous increase in the off-season unemployment rate During the on-season, however, there are offsetting forces on unemployment, with the sign of the net effect being unclear In addition, employment and the labor force increase Thus, even if unemployment increases in the on-season, the impact on the unemployment rate is ambiguous The unemployment rate is more likely to decrease the larger the expansion of on-season employment in response to SUA

Thus, the hypothesized effects of the introduction of an unemployment compensation program funded out of general revenues on the agricultural wage and salary worker unemployment rate and level of employment can be summarized as in table 1 [7] The off-season months for the

---

[7] The hypothesized effects on seasonality would be intensified by the response of unpaid family workers Since SUA is not experience rated, farm owners have an incentive to reclassify unpaid family workers as wage and salary workers and pay the appropriate

TABLE 1

HYPOTHESIZED EFFECTS OF SUA ON THE AGRICULTURAL WAGE AND SALARY LABOR FORCE

	Unemployment Rate (Seas Adj)	Level of Employment (Seas Adj)	Labor Force (Seas Adj)
Off-season	Increase	Decrease	Increase
On-season	Ambiguous	Increase	Increase
Annual (monthly average)	Increase	Ambiguous	Increase

TABLE 2

SEASONAL FACTORS FOR THE UNEMPLOYMENT RATE AND LEVEL OF EMPLOYMENT FOR AGRICULTURAL WAGE AND SALARY WORKERS

	Jan	Feb	March	April	May	June
Unemployment rate*	148 4	138 8	110 4	89 9	66 7	80 0
Employment	80 3	84 7	88 7	96 7	101 6	122 5
	July	Aug	Sept	Oct	Nov	Dec
Unemployment rate*	80 5	85 7	97 0	84 8	126 4	121 5
Employment	126 6	117 1	101 8	102 5	93 3	84 3

SOURCE —Bureau of Labor Statistics

* Implicit factor derived from dividing the unadjusted rate by the seasonally adjusted rate, 1974

unemployment rate and level of employment of agricultural wage and salary workers are November through April and the on-season months are May through October, with the exception that September appears to be an off-season month for the unemployment rate Seasonal factors for the employment status of agricultural wage and salary workers are shown in table 2

IV. Prediction Equations

The hypotheses are tested by comparing the observed seasonally adjusted unemployment rate and level of employment with values that we predict would have existed in the absence of the introduction of SUA in January 1975 The prediction equations are developed and estimated in this section The data on unemployment and employment used in this study are from the Current Population Survey (CPS) and refer to one's status in the reference week, the week containing the twelfth of the month

---

social security taxes to verify employment so as to collect benefits in the off-season This reclassification would increase the measured on-season employment of wage and salary workers but not of all farm-labor inputs and increase measured unemployment in the off-season However, because a reported earnings history is required for the receipt of benefits, unemployment increases from this factor are not likely to be important in the off-season months just following the enactment of SUA

A. *Unemployment Rate*

The prediction equation for the unemployment rate (URAG) uses seasonally adjusted quarterly data (average value for the 3 months) from 1948-I to 1974-III.[8] Quarterly data were used to reduce some of the noise that appears in the monthly CPS statistics

Agricultural wage and salary workers are predominantly male and disproportionately young[9] Seasonally adjusted unemployment in the agricultural sector would then be related to the stage of the nonfarm business cycle or the tightness in nonfarm labor markets for men with similar levels of skills It is for this reason that the two basic variables in the prediction equation are the seasonally adjusted unemployment rates of adult men (20 years of age and over) and of male teenagers (from 16 to 19 years of age) designated by URM20+ and URM16–19, respectively Time-trend variables (time and time squared) are also included to capture secular changes in agricultural unemployment rates relative to adult and teenage male unemployment rates[10]

With the unemployment rates expressed in natural logarithms, table 3 presents the regression equations Because of possible autocorrelation in the residuals of the ordinary least-squares equation (regression 1), it was recomputed with an adjustment for autocorrelation (regression 2)[11] The autoregressive parameter is RHO The Durbin-Watson statistic equal to 2 05 indicates that there is no serial correlation or seasonality in the residuals of the prediction equation

B *Level of Employment*

The equation for predicting the level of employment (EMAG) uses seasonally adjusted quarterly (average monthly) data from 1950-I to

---

[8] The prediction equations for the unemployment rate and the level of employment use data only up to 1974-III for two reasons First, it would be inappropriate to compare an observed value with a predicted value that is not independent of that observed value Second, because of autocorrelation in time-series data it is important to ascertain whether the observed value exceeds (or is less than) the predicted value in the period just prior to, as well as just after, the enactment of SUA

[9] In 1974, 83 percent of employed agricultural wage and salary workers were men, of whom nearly one-quarter were less than 20 years of age

[10] After controlling for the two unemployment rates and the curvilinear time trends, several alternative specifications of the minimum wage, which has covered some agricultural workers since 1967, were all statistically insignificant Since the objective was to obtain a good prediction equation rather than to analyze minimum-wage legislation, a minimum-wage variable was not included in the final equation For an analysis of the effect of minimum wages on farm labor markets, see Gardner (1972) The two demand-shift variables used in the analysis of employment (table 4) were also entered into the unemployment rate equation They had statistically insignificant effects separately and as a set and were therefore not included in the final regression

[11] For the data used in this paper (sample size approximately 100, four or five explanatory variables, 5-percent level of significance), the critical value of the Durbin-Watson statistic for rejecting the hypothesis of no autocorrelation is about 1 50 and for accepting the hypothesis of no autocorrelation about 1 70

TABLE 3

REGRESSION ANALYSIS OF AGRICULTURAL WAGE AND SALARY WORKER UNEMPLOYMENT RATES, 1948-I TO 1974-III

Explanatory Variables	OLS Regression	Adjusted for Autocorrelation
Intercept	− 0518	0587
	(−0 19)	(0 19)
LnURM20+	1634	2128
	(1 47)	(1 65)
LnURM16-19	6824	6058
	(3 93)	(3 04)
Time	0051	0057
	(2 45)	(2 18)
Time squared	− 00005	− 00005
	(−3 40)	(−2 83)
RHO		1747
		(1 75)
Adjusted R^2	73	73
N	107	106
D-W	1 67	2 05
SE	121	127

SOURCE —Bureau of Labor Statistics
NOTE —Seasonally adjusted quarterly data, t ratios in parentheses, dependent variable is LnURAG

1974-III It is a reduced-form equation in which employment is expressed as a function of the exogenous variables that shift the demand curve and the supply curve of agricultural wage and salary workers Since the objective is to predict employment in the absence of SUA, a reduced-form equation, rather than structural supply and demand equations, is sufficient

The demand for hired farm labor is a function of the price of the final good and the cost of alternative inputs The prices farmers receive are measured by a series produced by the Department of Agriculture (PR) On the assumption that farm owners purchase as nonlabor inputs a wide spectrum of industrial output in rough correspondence to the proportions produced for the entire economy, the cost of nonlabor inputs is measured by the wholesale price index for all industrial commodities (WPIIND) The partial effect of increased prices received increases the demand for farm labor Increased prices for industrial commodities would also increase the demand for farm labor if, as seems reasonable, the effect of the positive cross-elasticity of demand for labor exceeds the adverse scale effect on employment of higher production costs

The seasonally adjusted male teenage unemployment rate (URM16-19), which is dominated by job opportunities in the nonfarm sector, is used as the labor supply shift variable and is expected to have a positive effect on agricultural employment [12] Time-trend variables are included

---

[12] Although wage-rate data are available for the nonfarm sector, appropriate data do not exist for agricultural wage and salary workers

TABLE 4

REGRESSION ANALYSIS OF AGRICULTURAL WAGE AND SALARY WORKER EMPLOYMENT, 1950–I TO 1974–III

Explanatory Variables	Adjusted for Autocorrelation
Intercept	7 11
	(44 3)
LnPR	0 28
	(1 72)
LnWPI IND	0 51
	(1 71)
LnURM16–19	0 11
	(1 66)
Time	0 005
	(1 27)
Time squared	−0 0001
	(−2 60)
RHO	0 67
	(8 15)
Adjusted R^2	83
N	98
D-W	1 92
SE	0 053

SOURCE —Bureau of Labor Statistics and Department of Agriculture
NOTE —Seasonally adjusted quarterly data, t-ratios in parentheses, dependent variable is LnEMAG

in the prediction equation to capture the effects of other relevant but omitted demand and supply shift variables whose values may have changed over time

The regression, adjusted for autocorrelation, of agricultural wage and salary employment on the explanatory variables appears in table 4 [13] The Durbin-Watson statistic (1 92) implies that there is no serial correlation or seasonality in the residuals of the prediction equation

V. Testing the Hypotheses

Tables 5 and 6 compare the observed seasonally adjusted unemployment rate and level of employment with predicted values obtained from the regression equations in tables 3 and 4 adjusted for autocorrelation

The SUA was signed into law on December 31, 1974, with little forewarning [14] Since one's employment status in the CPS refers to the week containing the twelfth of the month, the January data should be

---

[13] Three other variables were entered and found to have F-ratios less than unity individually and as a set These were the unemployment rate of adult men, a farm minimum-wage variable, and the rate of change of nonfarm wages

[14] It is unlikely that the full response to a permanent SUA program could be observed in the first year since an employment history within the previous 12 months is required to receive benefits, changes in factor proportions require time, and the permanence of the program is uncertain

TABLE 5

UNEMPLOYMENT RATES OF AGRICULTURAL WAGE AND SALARY WORKERS (SEASONALLY ADJUSTED, %)

	NATURAL LOG OF THE UNEMPLOYMENT RATE			UNEMPLOYMENT RATE		
MONTH	Observed	Predicted*	Difference	Observed	Predicted†	Difference
1974:						
Oct	2 07	2 02	05	7 9	7 6	0 3
Nov	1 97	2 08	− 11	7 2	8 0	−0 8
Dec	2 07	2 10	− 03	7 9	8 2	−0 3
1975						
Jan	2 32	2 21	11	10 2	9 1	1 1
Feb	2 17	2 25	− 08	8 8	9 5	−0 7
Mar	2 48	2 25	23	12 0	9 5	2 5
Apr	2 53	2 34	19	12 6	10 4	2 2
May	2 24	2 34	− 10	9 4	10 4	−1 0
June	2 35	2 26	09	10 5	9 6	0 9
July	2 13	2 26	− 13	8 4	9 6	−1 2
Aug	2 35	2 26	19	10 5	9 6	0 9
Sept ‡	2 29	2 24	05	9 9	9 4	0 5

SOURCE —Observed data from Bureau of Labor Statistics

NOTE —Pre-SUA months are October 1974 through January 1975 off-season months after SUA are February through April 1975, on season months after SUA are May through September 1975

* Predicted from regression (2), table 3

† Antilog of the predicted value from table 3

‡ On the basis of seasonal factors September may be viewed as an off-season month for the unemployment rate

TABLE 6

EMPLOYMENT OF AGRICULTURAL WAGE AND SALARY WORKERS (SEASONALLY ADJUSTED)

	NATURAL LOG OF EMPLOYMENT (000)			EMPLOYMENT (000)		
MONTH	Observed	Predicted*	Difference	Observed	Predicted†	Difference
1974						
Oct	7 228	7 244	− 016	1,378	1,399	−21
Nov	7 234	7 227	007	1,386	1,376	10
Dec	7 148	7 215	− 069	1,272	1,360	−88
1975						
Jan	7 178	7 165	013	1,310	1,293	17
Feb	7 087	7 169	− 082	1,196	1,298	−102
Mar	7 085	7 111	− 026	1,194	1,225	−31
Apr	7 053	7 134	− 081	1,156	1,254	−98
May	7 203	7 109	094	1,344	1,223	121
June	7 115	7 195	− 080	1,230	1,333	−103
July	7 213	7 146	067	1,357	1,269	88
Aug	7 221	7 211	010	1,368	1,354	14
Sept	7 239	7 210	029	1,393	1,353	40

SOURCE —Observed data from Bureau of Labor Statistics

NOTE —Pre-SUA months are October 1974 through January 1975, off-season months after SUA are February through April 1975, on-season months after SUA are May through September 1975

* Predicted from the regression in table 4

† Antilog of the predicted value from table 4

viewed as a transition, as behavior is not likely to be altered significantly in 2 weeks. To the extent that there is more than a 6-week lag in response, perhaps February should also be viewed as a transition month. Three time periods will be compared: an off-season prior to SUA (October to December 1974), an off-season after SUA (February or March to April), and an on-season after SUA (May to September). Observed and predicted values are compared for October to December 1974 to test whether the prediction equation tends to consistently over- or underpredict for the months just prior to the start of SUA payments.

A. *Unemployment Rate*

During the base period (October to December 1974) the difference between the observed and predicted natural logarithm of the unemployment rate was very small in each month, with a 3-month average of −0.03, compared to a standard error of the prediction equation residual of .13. Hence, the equation does not tend to either overpredict or underpredict for months just prior to the enactment of SUA.

In the first off-season the mean difference between the observed and predicted natural logarithm of the unemployment rate was 0.11 for February to March and 0.21 for March and April. During the 4 on-season months, May to August, the difference was positive in 2 months, and the mean difference was −0.01. The difference equaled or exceeded 1.5 times the standard error of the residual for only 2 of the 12 months from October 1974 to September 1975 (the off-season months March and April), both of which have the hypothesized sign.

The mean difference between the observed and "predicted" (antilog of the predicted log value) unemployment rate was small for October to December 1974 and May to August 1975, −0.3 and −0.1 percentage points, respectively. For the off-season after the enactment of SUA it was 1.3 percentage points for February to April 1975 and 2.4 percentage points for March to April.

These data are consistent with the hypothesis that SUA increased the unemployment rate for agricultural wage and salary workers in the off-season but had no effect in the on-season. The off-season unemployment rate may have been increased by about 15 to 25 percent. This does not appear to be a recurring seasonal pattern, since the Durbin-Watson statistic indicates the absence of seasonality in the residuals of the prediction equation.

B. *Employment*

A similar analysis appears in table 6 for the level of employment. From October to December 1974 the mean difference between the observed and the predicted natural logarithm of employment (in thousands) was −0.026, half of the standard error of the residual (.053). As hypothesized,

TABLE 7

DIFFERENCE BETWEEN OBSERVED AND PREDICTED LABOR FORCE

	Months the Difference Is Positive	Average Monthly Difference (000)
Pre-SUA (1974)		
Oct.-Dec.	0 out of 3	−39
Post-SUA (1975)		
Feb.-Sept.	5 out of 8	−2
Mar.-Sept.	5 out of 7	+15

SOURCE.—Computed from tables 5 and 6.

the difference is consistently negative from February to April 1975, averaging −0.063 for the 3 months and −0.054 for March and April. The residual is positive in 4 of the 5 on-season months, May to September, averaging +0.024.

The hypothesized sign is obtained in 7 of the 8 months from February to September, or 6 of the 7 months from March to September. The probability of these outcomes (no more than one incorrect sign) appearing by chance is 3.5 percent and 5.3 percent, respectively. It could be asked, however, whether the signs of the differences between the observed and predicted values reflect a recurring seasonal pattern. If they did, the Durbin-Watson statistic would indicate a significant positive autocorrelation in the residuals of the prediction equation, but this is not the situation. In addition, the residuals were examined for the prior year (1974) and for 1958, the year with the highest post-World War II unemployment rate (6.8 percent) prior to 1975. In both years the hypothesized signs are obtained in 5 of the 8 months from February to September but not for the same 5 months. The probability of this outcome appearing by chance in each year is 36.3 percent.

Thus, the data are consistent with the hypothesis that SUA decreased employment during the off-season and increased employment during the on-season. There were approximately 65,000 to 70,000 fewer agricultural wage and salary jobs in the off-season months and approximately 31,000 additional jobs in the on-season months than would have been predicted.

C. *Labor Force*

Recall that in Section III it was hypothesized that SUA would increase the agricultural wage and salary worker labor force. This can be tested using the data in tables 5 and 6 to compute the observed and predicted labor force. The hypothesis is supported if the difference between the observed and predicted values is more positive (or less negative) after the enactment of SUA. This does, in fact, appear to be the situation (see table 7).

VI. Conclusions

The analysis in this paper suggests that the introduction in January 1975 of Special Unemployment Assistance (SUA), a nonexperience-rated federally funded unemployment compensation program, did affect the labor force status of agricultural wage and salary workers in the hypothesized manner

The unemployment rate during the off-season months was increased by approximately two percentage points, or roughly 20 percent The on-season unemployment rate was apparently unchanged Thus, the seasonality of the unemployment rate increased and the annual (average monthly) unemployment rate increased by approximately one percentage point (10 percent)

Employment in the off-season decreased by about 70,000 jobs per month (5 5 percent), while on-season employment increased by half that amount (31,000 additional jobs) Thus, the seasonality of employment increased and the annual level (monthly average) of employment of agricultural wage and salary workers appears to have decreased

Although separately the unemployment rate and employment analyses provide an affirmation of the hypothesis that SUA influenced labor market behavior, taken together the test is even stronger There does not appear to be any hypothesis, other than the rational response of employers and employees to SUA, that can explain the particular pattern of the data Yet, the quantitative findings may be underestimates of the longer-term effects on the hired farm labor force of a permanent SUA program, as there has been too little time for employers and employees to adjust fully to the new incentives

References

American Enterprise Institute *Unemployment Compensation Proposed Permanent Changes*. Washington: American Enterprise Inst , 1974

Council of Economic Advisers *Economic Report of the President, 1975* Washington Government Printing Office, 1975

Feldstein, Martin "Temporary Layoffs in the Theory of Unemployment" *J P E* (1976), in press

Gardner, Bruce "Minimum Wages and the Farm Labor Market" *American J Agricultural Econ* 54 (August 1972) 473–76.

O'Connor, James "Seasonal Unemployment and Unemployment Insurance" *A E R* 52 (June 1962) 460–71

U S Department of Labor *Comparison of State Unemployment Insurance Laws, January 1972* Washington Government Printing Office, 1972

———. *Significant Provisions of State Unemployment Insurance Laws, July 15, 1975* Washington. Manpower Admin , Unemployment Insurance Service, 1975.

A Dynamic Model of the Firm

Arvind Virmani

State University of New York at Albany

I. Introduction

The usual neoclassical production function forms the point of departure for the present model of the firm. I will therefore briefly review the conceptual basis of capital and of the production function in the neoclassical model of the firm.[1] In the static model, maximization of profits under the competitive assumption yields the usual optimality conditions that the firm must equate marginal value product of capital to its rental value and the marginal value product of labor to the wage rate. Under decreasing returns, this yields the optimal level of capital and labor used by the firm, while under constant returns to scale in production only the capital-labor ratio can be determined, the size of the firm being indeterminate.

There is little difference between the conception of capital implicit in the production function and the way land may be treated in an agricultural production function. However, it is possible to modify the analysis by a consideration of the short- and long-run and the associated fixed and variable factors and thus introduce features which distinguish capital from land. Thus, variable factors like labor are adjustable in the short run but not the fixed factors such as capital. The latter are adjustable over the long run.

In the intertemporal version of the model,[2] the firm maximizes the discounted stream of returns. Its capital stock is obtained through investment. This treatment gives an explicit relation between the rental value of capital and the cost of investment ($[r + \delta - \dot{c}/c]c$) as long as there is a positive level of new investment (by any firm in the economy). If the aggregate demand for this good is negative, the rental value of capital must adjust to equate supply and demand for existing capital. A paradoxical result of this model is that capital again becomes a variable factor. Also, for discontinuous changes in the prices of output or inputs, the quantity of investment is indeterminate. In this version of the model,

---

[1] By the "neoclassical model" I mean the textbook theory which forms the basis of most of the precise predictions and analysis in microtheory.

[2] We are concerned with the formulation due to Jorgenson (1967).

[*Journal of Political Economy*, 1976, vol. 84, no. 3]

© 1976 by The University of Chicago. All rights reserved.

the view of capital is modified somewhat from the previous one. Capital can here be conceived of as consisting of putty (or Swan's more prosaic concept of a meccano set) which one can mold, add to, or subtract from as desired.

There are two major criticisms of the model outlined. These are the assumption of homogeneity and intertemporal substitutability of capital. The former problem applies equally to labor inputs and is not a serious one as far as the theory of the firm is concerned.[3] In principle, one can incorporate different types of capital and labor into a production function. The intertemporal substitutability of capital is a more fundamental problem. If capital exists in the form of putty, there is no problem of adding up the capital bought in different periods or using it interchangeably in production. (The same is true of Swan's identical meccano sets with changing technical recipes.) In fact, capital consists of machines, equipment, plants, etc., which undergo technical change.[4] Furthermore, productive facilities exist in the form of plants which have an optimal or planned level of labor use which cannot be freely varied.[5] For both these reasons, the assumption of intertemporal substitutability of capital in production is a highly dubious assumption for much of the industrial sector. It is less inappropriate for agriculture and service sectors.

The focus of my analysis is therefore the plant or production unit. At the simplest level, we can think of this as being constructed from a homogeneous investment good.[6] The range of choices from among which a firm can select a plant for construction is represented by the plant-choice function. This represents the available choice of techniques or capital-labor ratios, the corresponding sizes of plants, and the output available from each such plant. Once the techniques have been chosen and incorporated into a plant they must remain fixed for the life of the plant. For the purpose of this paper, this involves keeping the chosen capital-labor ratio constant.[7] From a theoretical standpoint, the change to note is the replacement of the production function by the plant-choice function. This provides a solution to the most important criticism of the former by obviating the need for intertemporal aggregation of capital.[8]

Technical change is theoretically classified into two kinds—embodied and disembodied. The former is represented here as improvements in the investment good and in the productivity of the plant as a unit. At this homogeneous-good level, the latter can be thought of as represented in the capital-labor organization of the plant. Disembodied technical change can be incorporated into existing plants. This type of technical change can be subdivided into an exogenous and an endogenous component. The former is represented by the exogenous increase in the educational level of the labor pool, which may be transferred to the plant through the usual process of labor turnover. The latter is a kind of learning by doing arising from increased familiarization with and adaptation to the new plant on the part of labor.

---

[3] For a review of the problems centering on this question in aggregate growth theory, see Harcourt (1972) and Stiglitz (1974).

[4] Solow (1959) was the first to consider embodied technical change in the context of an aggregate growth model.

[5] Johansen (1959) was the first to consider ex ante substitutable ex post fixed coefficients, again in the context of aggregate growth theory.

[6] In a subsequent paper, I will consider the inhomogenous-good case.

[7] A more general model which permits adjustments in this ratio will be introduced in a subsequent paper.

[8] This does not exclude the possibility of such aggregation of certain types of capital in a nonhomogeneous-goods framework, as will be shown in a subsequent paper.

II. The Basic Model

A firm consists of a number of plants or production units. We will assume for the sake of simplicity that the firm produces a single product (it is easy to generalize to a multiproduct firm). Therefore, in this case each plant produces the same product. This output is sold on a perfectly competitive market. The operational significance of this assumption is that there are no external limitations on how much the firm can sell at the given market price. The firm's objective is to maximize the present value of its profit stream, given by

$$\int_0^\infty A(t)\;[p\;Q(t)\;-\;w\;L(t)\;-\;c\;I_t]\;dt, \tag{1}$$

where $Q(t)$, $L(t)$, and I_t are the total output produced, total labor employed, and the investment, respectively, at time t. p, w, and c are the prices of output, labor, and investment goods, respectively, at time t, and $A(t)$ is the discount factor. In general, this may be related to the ease with which funds can be obtained and the returns on borrowed or lent funds. If perfect financial markets are assumed,

$$A(t) = e^{-rt}, \tag{1'}$$

where

$$r(t) = \int_0^t R(v)\;dv,$$

and $R(v)$ is the instantaneous rate of interest at time v.

Each period the firm must decide on the investment to be made in new plants. As I noted in Section I, the plant-choice function represents the range of techniques and sizes of plants from which the firm can choose at any given point in time. We will assume that this function is a continuous twice-differentiable function of investment and labor. Or, $Q_v = F_v(I_v, L_v)$. Thus, investment at any point of time v involves choosing the level of investment I_v and the capital-labor ratio $k_v = I_v/L_v$ for the new plant Once the plant is built, this ratio is assumed to remain fixed.[9] The planned or initial labor usage in the plant is L_v, and Q_v is the planned or initial output of this plant As the subscript on the F indicates, the plant-choice function can change with time with the development of new techniques and technology We will assume that this function satisfies the following conditions (with the numerical subscripts representing differentials with respect to the first or second arguments) $F_{v1}, F_{v2}, F_{v12}, F_{v21} \geq 0$, $F_{v11}, F_{v22} \leq 0$, $J = F_{v11} \ F_{v22} - F_{v12} \ F_{v21} \geq 0$ The last condition, of course, means that the plant-choice function must have nonincreasing returns to scale We will also assume that the plant-choice function is homogeneous of degree h

Depreciation is assumed to manifest itself as the decay of capital at a constant exponential rate, δ The stock of capital at time t in the plant of vintage v is $I_v(t) = I_v \ e^{-\delta \ (t-v)}$ To maintain the constancy of the capital-labor ratio, it follows that the labor employed in the plant vintage v at time t is $L_v(t) = L_v \ e^{-\delta(t-v)}$ It is possible to think of capital decay as a decrease in the effective capital stock due to depreciation, which permits a proportional decrease in the labor used in the plant Disembodied technical change is assumed to manifest itself as an increase in output over time[10] at a constant exponential rate, σ As was indicated in Section I, the endogenous part of this is a type of learning by doing This is a kind of learning by doing which depends only on the time elapsed since the plant was built and staffed In effect, any learning prior to the start of the new plant is already included in the vintage-dependent term for general plant technology [11]

The output from a plant of vintage v is therefore influenced by both the depreciation and the disembodied technical change It is therefore given by

$$\begin{aligned} Q_v(t) &= e^{\sigma(t-v)}F_v[I_v(t), L_v(t)] = e^{\sigma(t-t)}F_v[I_v e^{-\delta(t-v)}, L_v e^{-\delta(t-v)}] \\ &= e^{-\delta \, (t-v)}F_v(I_v, L_t), \end{aligned}$$

where $\delta' = h \ \delta - \sigma$ The total output of the firm from all plants is therefore

$$Q(t) = \int_{-\infty}^{t} Q_v(t) \, dv = \int_{-\infty}^{t} F_v(I_v, L_v) \ e^{-\delta \, (t-v)} \, dv \tag{2}$$

---

[9] A certain type of fixed change in this ratio is permitted in another paper A more general model which permits freer adjustment is not included in this paper but will be presented in a subsequent paper

[10] The only other possibility is a decrease in labor requirements over time It is conceptually more satisfactory to introduce this along with the more realistic depreciation model

[11] Note that the usual representation of learning by doing as a function of the cumulative output reduces to one which is similar to that used here if the output level is kept constant

On differentiation this yields the following constraint

$$\dot{Q}(t) = F_t(I_t, L_t) - \delta' \int_{-\infty}^{t} F_v(I_v, L_v) \; e^{-\delta(t-v)} \, dv$$

or

$$\dot{Q}(t) = F_t(I_t, L_t) - \delta' \; Q(t), \tag{3}$$

where $\dot{Q}(t) = dQ/dt$ The total labor used by the firm is similarly

$$L(t) = \int_{-\infty}^{t} L_v(t) \, dv = \int_{-\infty}^{t} L_v \; e^{-\delta(t-v)} \, dv$$

or

$$\dot{L}(t) = L_t - \delta \; L(t), \tag{4}$$

where $\dot{L}(t) = dL/dt$

The constraint equation (3) says that the change in output at time t must equal the output due to investment at time t, minus the output decrease due to decay of capital, plus the output increase due to the increase in the productivity of old plants[12] The constraint equation (4) says that the increase in labor employed by the firm is given by the labor needed for the new plants minus the labor released from the old plants due to capital decay

The problem is therefore to maximize (1) subject to the constraints (3) and (4) The necessary conditions for an optimum, which are derived in the Appendix, after simplification and use of equation (1) yield the following equations[13]

$$F_{t1}(I_t, L_t) = c/P_s, \tag{5}$$

$$F_{t2}(I_t, L_t) = W_s/P_s, \tag{6}$$

$$P_s = e^{mt} \int_{t}^{\infty} p \; e^{-mt} \, dt, \qquad m = r + \delta', \tag{7}$$

$$W_s = e^{nt} \int_{t}^{\infty} u \; e^{-nt} \, dv, \qquad n = r + \delta, \tag{8}$$

where F_{t1} and F_{t2} are the differentials of the function F with respect to its first and second arguments, I_t and L_t, respectively The former represents the marginal product of investment in a new plant while the latter is the marginal product of planned (or initial) labor inputs into the new plant Equations (7) and (5) indicate that P_s can be termed the shadow price of the output produced by the new plant The former equation shows that it is related to the stream of output produced by this plant, appropriately adjusted for continuing depreciation and technical change Thus equation (5) shows that investment in a new plant must be carried to the point at which the shadow marginal value of output obtained from it is equal to the cost of investment Similarly, equations (8) and (6) indicate that W_s is the shadow wage of labor Again, the former equation indicates how it represents the stream of labor inputs into the plant over its life, adjusted for the depreciation-related release of labor Equation (6) then shows that the planned labor input into the plant must be such that its shadow marginal value product equals its shadow wage

---

[12] Recall that $\delta' = \delta \; h - \sigma$ The first term on the right represents the effects of depreciation and the second term the effect of continuing technical change

[13] The sufficient conditions for a maximum are also examined in the Appendix The earlier assumption of nonincreasing returns is related to these conditions

It is clear from the above equations that if future prices and wages and the discount rate are known, P_s and W_s can be calculated These can then be substituted into equations (5) and (6) to obtain the optimum level of investment and the capital-labor ratio for the new plant (If $J = 0$ only the latter can be determined) We thus obtain optimal investment as a function of the present and future price of output and labor, the present and future rate of interest, and the present price of investment goods The total output of the firm is obtained from equation (3) and is also a function of the same variables The important role that future prices and wage rates play in the present decisions is apparent from the optimality conditions The present price and wage rate constitute only a small element of the shadow price of output and the shadow wage, respectively

If the exogenous variables mentioned above are given to the firm through a futures market, the firm has no need to know either the future price of investment goods[14] or the future shape of the plant-choice function as far as the present decision is concerned The latter means that the firm does not need to be concerned about future technical innovations However, it is implicit in the perfect competition assumption that both these are already represented in the future price of output The mechanism of action is the future supply curve of the industry to which this firm belongs Thus a future decrease in the price of investment goods or a positive increase in the rate of technological change shifts the supply curve to the right, reducing the output price at the time (This statement implicitly makes use of the analysis of Section III) This decrease in price would be reflected in the future market for output and thus influence the present shadow price and the present investment decision [15]

III. Analysis

An explicit solution of equations (5) and (6) would, of course, require the assumption of a particular functional form for F_t We can, however, analyze the effect of changes in exogenous variables on investment, labor use, etc, before making this choice Totally differentiating (5) and (6), for given t, we obtain[16] $F_{11}\ dI_t + F_{12}\ dL_t = dc/P_s - c\ dP_s/P_s^2$, $F_{21}\ dI_t + F_{22}\ dL_t = dW_s/P_s - W_s\ dP_s/P_s^2$ These are easily solved (if $J = 0$) to get

---

[14] This is no longer so if investment in a given plant is distributed over time

[15] Note that participation in the futures market would require that the firm determine its future demands and supplies

$$\begin{bmatrix} dI_t \\ dL_t \end{bmatrix} = \frac{1}{J} \begin{bmatrix} F_{22} & -F_{12} \\ -F_{21} & F_{11} \end{bmatrix} \begin{bmatrix} 1/P_s & 0 & -c/P_s^2 \\ 0 & 1/P_s & -W_s/P_s^2 \end{bmatrix} \begin{bmatrix} dc \\ dW_s \\ dP_s \end{bmatrix}$$

Using equations (5) and (6) to simplify we obtain

$$dI_t = F_1\ [F_{22}\ dc - F_{12}\ dW_s + (F_2F_{12} - F_1F_{22})\ dP_s]/(cJ) \qquad (9)$$

$$dL_t = F_1\ [-F_{21}\ dc + F_{11}\ dW_s + (F_1F_{21} - F_2F_{11})\ dP_s]/(cJ) \qquad (10)$$

Similarly, differentiating (7) and (8) and simplifying[17] we get

$$dP_s = dp/m - \left(e^{mt} \int_t^{\infty} P_s\ e^{-mv}\ dv\right)\ dr \qquad (11)$$

$$dW_s = dw/n - \left(e^{nt} \int_t^{\infty} W_s\ e^{-nv}\ dv\right)\ dr \qquad (12)$$

Here the differentials with respect to p or w represent differentials with respect to the whole price or wage rate stream, respectively That is, it represents changes which occur when the whole price or wage rate stream is shifted marginally It is easy to show, however, that the signs obtained remain the same if the changes are for a given time interval or for a stream starting at $t_0 > t$ In the latter case we have

$$\partial P_s/\partial p = e^{-m(t_0 - t)}/m \qquad (11')$$

$$\partial W_s/\partial w = e^{-n(t_0 - t)}/n \qquad (12')$$

This shows that the further in the future the changes commence (i e , the greater t_0 is), the smaller their effect on present (i e , at $t = 0$) shadow price or shadow wage rate

Using the differential equations derived above and the assumptions on the plant-choice function presented in Section II, we can carry out the following analysis We have $\partial I_t/\partial c = F_1F_{22}/cJ < 0$, $\partial L_t/\partial c < 0$ Both investment and the planned labor input into the new plant are negatively related to the present cost of investment goods From equation (2), noting that investment and labor use in old plants are unaffected by present or future changes, we must have

$$\begin{aligned} \partial Q/\partial x &= \int_{t-dt}^{t} e^{-\delta(t-v)}\ [F_1(\partial I_t/\partial x) + F_2(\partial L_t/\partial x)]\ dv \\ &= (1 - e^{-\delta}\ dt)\ [F_1(\partial I_t/\partial x) + F_2(\partial L_t/\partial x)]/\delta', \end{aligned} \qquad (13)$$

---

[16] I drop the t subscript on F for the rest of this paper

[17] See Appendix We are again holding t constant

where x is c, w, or p As $\delta' dt > 0$ and $e^{-\delta' dt} < 1$, $\partial Q/\partial c < 0$ The total output of the firm is reduced if the price of investment goods rises.

Similarly, we have $\partial I_t/\partial w = -F_1F_{12}/nJc < 0$, $\partial L_t/\partial w < 0$, $\partial Q/\partial w < 0$ This shows that changes in the present or future wage rate have a negative effect on the investment and planned labor input into the new plant as well as on the total output of the firm Using equation (12′) we can also see that the further in the future the exogenous perturbances in the wage rate appear, the smaller will be the absolute effect on these variables We also have $\partial I_t/\partial p = F_1(F_2F_{12} - F_1F_{22})/mJc > 0$, $\partial L_t/\partial p > 0$, $\partial Q/\partial p > 0$ Thus, changes in the present or future prices of output have a positive effect on investment and planned labor input in the new plant as well as on total output, and the nearer to the present these changes start the greater is the effect

For the effect of a change in the capital-labor ratio, $k_t = I_t/L_t$, we have[18] $(\partial k_t/\partial c)/k_t = (\partial I_t/\partial c)/I_t - (\partial L_t/\partial c)L_t = F_1(F_{22}/I_t + F_{21}/L_t)/cJ = F_1(F_{21}I_t + F_{22}L_t)/(cJI_tL_t)$ Now as F is homogeneous of degree h less than one, F_2 is homogeneous of degree $h - 1 < 0$ Therefore, $F_{21}I_t + F_{22}L_t = F_2(h - 1) < 0$, and $\partial K_t/\partial c < 0$ We can similarly show that $\partial k_t/\partial w_t = -k_tF_1^2(h - 1)/(ncJI_tL_t) > 0$, and $\partial k_t/\partial p = 0$

From equations (9), (11), and (12) we have

$$\begin{aligned}\partial I_t/\partial r &= \left[(F_1F_{22} - F_2F_{12})\ e^{mt}\int_t^\infty P_se^{-mv}\,dv + F_{12}\ e^{nt}\right.\\ &\quad\left.\int_t^\infty W_se^{-nv}\,dv\right]\cdot F_1/cJ\\ &= -\left(\partial I_t/\partial p\ \ me^{mt}\int_t^\infty P_se^{-mv}\,dv + \partial I_t/\partial w\ \ ne^{nt}\right.\\ &\quad\left.\int_t^\infty W_se^{-mv}\,dv\right)\end{aligned} \tag{14}$$

We know that both

$$\int_t^\infty P_se^{-mv}\,dv \quad\text{and}\quad \int_t^\infty W_s\ e^{-nv}\,dv$$

are positive As $\partial I_t/\partial p$ and $\partial I_t/\partial w$ have opposing signs, the sign of (14) depends on whether the first or the second term on the right is greater in absolute value We can call the first term the price effect and the second one the wage effect If the first is stronger than the second, then the effect of discount-rate changes on investment in the new plant is negative, but otherwise it is positive This is a somewhat unexpected result An examination of equations (11) and (12) indicates how the price and wage effect terms arise These show that an increase in the discount rate has the effect of decreasing the shadow price and the shadow wage rate This is the inverse of the effect that an increase in p and w has on these two, respectively Thus, an increase in the rate of discount is similar in sign to a simultaneous decrease in the price and wage rate after adjustment is made for the different weighting factors A similar analysis is applicable to the effect of discount-rate changes on the planned labor use and the total output of the firm, both being ambiguous in sign.

---

[18] My thanks to Professor Arrow for suggesting the methodology for determining the signs

This analysis also raises the possibility of investment (and the other dependent variables) being independent of the discount rate It seems unlikely that the two effects would balance each other exactly However, it is possible that the net effect lies in a narrow band around zero In such a case, fluctuations in the exogenous variables could cause it to fluctuate around zero Any observations based on average behavior over time would then indicate a zero net effect

IV. Conclusions

There are two aspects of this model which distinguish it from the usual neoclassical model One relates to the assumptions on which it is based and the other to the results derived from it The former have been reviewed in Section I and their correspondence to reality emphasized there Among the latter, the importance of future prices and wage rates was indicated in Section II, and the possibility of a positive or near zero effect of interest-rate changes on investment was shown in Section III Another implication of this model is the change in focus by the firm, from the passive adjustments which characterize the neoclassical model to active pursuit of investment and growth In the neoclassical model the firm reacts to price changes by appropriate adjustments in capital and labor inputs so as to maintain optimality In my model the perspective is changed considerably Adjustments take place in the level of investment and the capital-labor ratio to be used in the new plant

Appendix

The problem is to maximize (1) subject to (3) and (4) We form the Hamiltonian H, $H = A(t)(pQ - wL - cI_t) + P'_s\,[F_t(I_t, L_t) - \delta' Q] + W'_s\,(L_t - \delta\ L)$ The necessary conditions are then given by[19] $\partial H/\partial I_t = -cA(t) + P'_s\ \ F_{t1}(I_t, L_t) = 0$, or

$$\begin{aligned} F_{t1}(I_t, L_t) &= c/P_s, \qquad P_s = P'_s = P'_s/A(t) \\ \partial H/\partial L_t &= P'_s\ \ F_{t2}(I_t, L_t) + W'_s = 0, \end{aligned} \tag{15}$$

---

[19] See, for example, Bryson and Ho (1969)

or

$$F_{t2}(I_t, L_t) = W_s/P_s, \qquad W_s = -W_s'/A(t)$$
$$\dot{P}_s' = -\partial H/\partial Q = -[pA(t) - \delta' P_s'], \tag{16}$$

or

$$\dot{P}_s' - \delta' P_s' = -PA(t) \quad \text{or} \quad d(P_s' e^{-\delta' t})/dt = -pA(t)e^{-\delta' t}$$

or

$$P_s' = e^{\delta' t} \int_t^{\infty} pA(t)e^{-\delta' v}\, dv \tag{17}$$

The $W_s' = -\partial H/\partial L = -[-wA(t) - \delta W_s']$ or, solving as above,

$$W_s' = -e^{-\delta' t} \int_t^{\infty} wA(t)e^{-\delta v}\, dv \tag{18}$$

Equations (15)–(18) reduce to (5)–(8) if we substitute $A(t) = e^{-rt}$ from (1′) and put $m = r + \delta'$, $n = r + \delta$

The concavity assumptions on F_t along with the linearity of the rest of H mean that the concavity of H^0 is insured (Arrow and Kurz 1970) In addition, a maximum requires the fulfillment of the following transversality conditions

$$\lim_{t \to +\infty} P_s' \geq 0, \qquad \lim_{t \to +\infty} P_s' Q = 0,$$
$$\lim_{t \to +\infty} W_s' \geq 0, \qquad \lim_{t \to +\infty} W_s' L = 0$$

Let us consider these conditions for the case in which r is constant and p and w are growing at the constant rates a and b, respectively Substituting $p = p_0 e^{at}$, $w = w_0 e^{bt}$, and $A(t) = e^{-rt}$ in equations (17) and (18), we obtain $P_s' = p_0 e^{(a-r)t}/(m - a)$, $W_s' = w_0 e^{(b-r)t}/(n - b)$ For P_s', the transversality condition therefore requires that r be greater than a This is a very plausible condition because if the reverse were true, it would pay to keep goods off the market and keep accumulating them for future sales A similar condition must be fulfilled for W_s' For the more general case in which p and w are not growing at a constant rate, we would expect such a condition to be fulfilled on the average, though at any individual point it may be violated For the constant-growth case, the second set of transversality conditions requires that $\dot{Q}/Q = Q_t/Q - \delta' \leq r - a$, $\dot{L}/L = L_t/L - \delta \leq r - b$ The first equation says that the rate of increase in the production of output must be less than or equal to $(r - a)$ or that the rate of growth of the value of output be less than the discount rate The second equation is a similar statement about wage outlays Both these are linked to the convergence of the maximand in (1) In this and most subsequent analyses we will assume that the transversality conditions are fulfilled

Differentiating equation (7) we obtain

$$\frac{\partial P_s}{\partial p} = e^{mt} \int_t^{\infty} e^{-mv}\, dv = \frac{e^{mt} e^{-mt}}{m} = \frac{1}{m}$$

$$\frac{\partial P_s}{\partial r} = te^{mt} \int_t^{\infty} pe^{-mv}\, dv - e^{mt} \int_t^{\infty} tpe^{-mv}\, dv$$

$$= tP_s + e^{mt} \int_t^{\infty} tpe^{-mv}\, dv$$

Integrating the second term by parts and substituting P_s, we have

$$\frac{\partial P_s}{\partial r} = tP_s + e^{mt}\left(t\int_t^{\infty} pe^{-mv}\ dv - \int_t^{\infty}\int_t^{\infty} pe^{-mv}\cdot dv\right)$$

$$= tP_s - tP_s - e^{mt}\int_t^{\infty} P_s e^{-mv}\ dv$$

$$= -e^{mt}\int_t^{\infty} P_s e^{-mv}\ dv$$

The analogous differentials of W_s are derived from equation (8) in exactly the same way.

References

Arrow, K. J., and Kurz, M. *Public Investment, the Rate of Return, and Optimal Fiscal Policy.* Baltimore: Johns Hopkins Press (for Resources for the Future), 1970.

Bryson, A. E., and Ho, Y. *Applied Optimal Control.* Boston: Ginn, 1969.

Harcourt, G. C. *Some Cambridge Controversies in the Theory of Capital.* London: Cambridge Univ. Press, 1972.

Johansen, L. "Substitution versus Fixed Proportions in the Theory of Economic Growth: A Synthesis." *Econometrica* 27 (April 1959): 157–76.

Jorgenson, D. W. "The Theory of Investment Behavior." In *Determinants of Investment Behavior, a Conference of the Universities-National Bureau Committee for Economic Research*, edited by R. Ferber. New York: Nat. Bur. Econ. Res., 1967.

Solow, R. M. "Investment and Technical Progress." In *Stanford Symposium on Mathematical Methods in the Social Sciences*, edited by K. J. Arrow, S. Karlin, and P. Suppes. Stanford, Calif.: Stanford Univ. Press, 1959.

Stiglitz, J. E. "The Cambridge-Cambridge Controversy in the Theory of Capital; A View from New Haven: A Review Article." *J.P.E.* 82, no. 4 (July/August 1974): 893–903.

Multilateral and Bilateral Measures of Effective Exchange Rates in a World Model of Traded Goods

Stanley W. Black

Vanderbilt University and Institute for International Economic Studies, University of Stockholm

When the exchange rates of most of the major trading nations of the world are floating, the two-country theory of the balance of payments with traded and nontraded goods needs to be modified to take account of the multiplicity of foreign price levels for tradable goods.[1] This note develops the relationship between the price of traded goods in a single country and changes in exchange rates in all other countries, using a simplified version of the IMF's Multilateral Exchange Rate Model (MERM) (see Artus and Rhomberg 1973). Some notable simplifications are obtained under small-country assumptions.

The standard technique for dealing with a group of floating rates is the concept of *effective exchange rate* (see *Economic Report of the President* 1974, pp. 220–26). When there are n countries floating, the movement of the dollar exchange rate may give a very inaccurate idea of real changes in the international purchasing power of any given currency. The effective exchange rate for a country can be defined as a weighted average of the exchange rates of its trading partners, with all rates being measured relative to some base year. An index can then be computed by comparing the actual exchange rate of the country with the weighted average of the rates of its trading partners.

The simplest type of weighting system is a bilateral system of weights, using either export shares or import shares or an average of both. For example, an export-weighted effective rate index can be defined as in equation (1), where

$$e_j = \sum_{k \neq j} E_{jk} r_k, \qquad \text{index} = r_j / e_j \tag{1}$$

the rates r_k are defined as units of currency k per dollar, and $E_{jk} = X_{jk}/\sum_{k \neq j} X_{jk}$ = share of j's total exports going to k where X_{jk} is the flow

---

[1] The theory is given for a two-country world by Kemp (1970).

[*Journal of Political Economy*, 1976, vol. 84, no. 3]
© 1976 by The University of Chicago. All rights reserved.

from j to k. Similarly, an import-weighted index is defined in (2), where $M_{kj} = X_{kj}/\sum_{k \neq j} X_{kj}$ = share of j's imports coming from k:

$$e_j = \sum_{k \neq j} M_{kj} r_k, \qquad \text{index} = r_j/e_j. \tag{2}$$

These indexes give extremely useful information concerning the effect of shifts in world exchange rates on the competitive position of a given country. Nevertheless, they have at least two major shortcomings. One generally acknowledged drawback is that bilateral weight indexes ignore the effects of competition between the products of countries j and k in third markets. An additional problem is that an index of exchange rates is not a price index of tradable goods. Both of these limitations can in principle be resolved through the use of the MERM.

1. Prices of Traded Goods

The following simplified version of MERM assumes that there is only one class of tradable good. It is also assumed, for simplicity, that incomes and prices of nontraded goods are held constant by policy in all countries while exchange rates are allowed to fluctuate in response to exogenous shocks. For any country k, the tradable "good" Z^k is assumed to be a composite function of the "products" of the n trading partners

$$Z^k = f^k(Z_1^k, \ldots, Z_n^k) \tag{3}$$

In order for this aggregate tradable good to be definable independently of the amount of the nontradable good, it must be assumed that the marginal rate of substitution between any two "products" in (3) is independent of the consumption of the nontradable good (see Armington 1969).

Under these assumptions, the proportionate change in the demand for exports from country j to country k can be written as

$$Z_j^{*k} = \sum_{i=1}^{n} \eta_{j/i}^{k}(P_i^* - T_k^*), \tag{4}$$

where P_i^* = proportionate change in the dollar price of the product of country i, T_k^* = proportionate change in the dollar price of country k's currency, and $\eta_{j/i}^k$ = price elasticity of demand in country k for product j with respect to price i.[2] With $X_j^k = P_j Z_j^k$ defined as the dollar value of exports from j to k, the share of j's total product of tradable goods exported to k is $W_j^k = X_j^k/\sum_{k=1}^{n} X_j^k$. The percentage change in the total value of production of tradables in j is then

$$X_j^* = \sum_k W_j^k X_j^{*k} = P_j^* + \sum_k W_j^k Z_j^{*k}.^{3} \tag{5}$$

---

[2] Artus and Rhomberg (1973, eq. [1], simplified).

[3] Ibid., eq. (3), simplified.

Finally, we assume that the price elasticity of supply of tradables **produced in country j is β_j, so** that the proportionate change in the value **of tradables in country j is**

$$P_j^* + \beta_j(P_j^* - T_j^*)^{4} \tag{6}$$

Equating value of demand and supply for each country j, we have the equilibrium conditions

$$\sum_{k=1}^{n} W_j^k \sum_{i=1}^{n} \eta_{j/i}^k (P_i^* - T_k^*) = \beta_j(P_j^* - T_j^*), \qquad j = 1, \ldots, n \tag{7}$$

The demand equations have the property that

$$\sum_{i=1}^{n} \eta_{ji}^k = \eta^k, \tag{8}$$

since an equal proportionate increase in the tradable-goods prices of all countries will reduce demand in country k depending on the demand elasticity in k After dividing through by β_j, we can write (7) as

$$\sum_{i=1}^{n} \left(\frac{1}{\beta_j} \sum_{k=1}^{n} W_j^k \eta_{ji}^k - \delta_{ij} \right) P_i^* = \sum_{k=1}^{n} \left(\frac{1}{\beta_j} W_j^k \eta^k - \delta_{jk} \right) T_k^*, \tag{9}$$

where $\delta_{ij} = 1$ if $i = j$ and zero otherwise, and similarly for δ_{jk} In matrix terms, (9) can be written as

$$Ap^* = Bt^*, \tag{10}$$

with solution

$$p^* = A^{-1}Bt^* = Mt^* \tag{11}$$

Note that if all $\beta_j = \infty$, $A = B = -I$ Rewriting (11), we have

$$P_j^* = \sum_{k=1}^{n} m_{jk} T_k^*, \tag{12}$$

where the m_{jk} are the elements of a matrix depending on the trade shares and the demand and supply elasticities The implication of this result is that the changes in exchange rates T_k^* now have definite implications for the changes in prices of tradable goods produced in each country If the supply elasticities β_j are infinite, the m_{jk} of (12) equal unity for $j = k$ and zero otherwise, implying that $P_j^* = T_j^*$ for all j (This is obvious from [7], after both sides are divided by β_j)

Finally, we need to relate the prices (P_j) of the "products" (Z_j^i) of each country j to the price (P^i) of the tradable "good" (Z^i) consumed in country i This can be done by means of the following price index Let $S_j^i = X_j^i / \sum_{j=1}^{n} X_j^i$ = the share of i's total use of the tradable good

---

[4] Ibid, eq (15), simplified, with the supply elasticity β_j defined in volume terms

that is imported from j. Then Armington (1969, p. 174, n. 29) shows that a price index for the tradable good can be defined as

$$P^{*i} = \sum_{j=1}^{n} S_j^i P_j^* \tag{13}$$

Making use of (12), we have a relation between the changes in exchange rates in all countries and the world price of tradable goods in country i.

$$P^{*i} = \sum_{j=1}^{n} S_j^i \sum_{k=1}^{n} m_{jk} T_k^* = \sum_{k=1}^{n} \left(\sum_{j=1}^{n} S_j^i m_{jk} \right) T_k^* \tag{14}$$

When supply elasticities are infinite, this simplifies to

$$P^{*i} = \sum_{k=1}^{n} S_k^i T_k^*, \tag{15}$$

which is very close to the bilateral import-weighted index given by equation (2). In fact, if we define Z^k to be the foreign-source tradable good, (15) would be exactly the same as (2) except that the exchange rates are defined inversely. Thus, computations based on formula (2) are consistent with MERM and will be useful measures of the effects of exchange rates on import prices for small countries facing fixed prices.

A further question is, Under what conditions may formula (1), the export-weighted index, be consistent with MERM? There are three sets of conditions in which a formula like (1) can be a useful measure of export prices. First, if the country is a small country unable to affect the prices of its exports, then the elasticity of demand for its product $\eta_{j/j}^k$ tends to minus infinity. If we assume that the elasticity of substitution between the different "products" in (3) also tends to infinity, then the prices of all the related goods move together, and other cross-elasticities go to zero. In this case the export price of a group of exporters of a homogeneous product is determined by the following modification of equation (7) for the case of infinite demand elasticities

$$\sum_{k \neq j} W_j^k P_j^* = \sum_{k \neq j} W_j^k T_k^* \tag{16}$$

or

$$P_j^* = \sum_{k \neq j} E_{jk} T_k^*, \tag{17}$$

where the weights $E_{jk} = W_j^k / \sum_{k \neq j} W_j^k$ are defined for a group of exporters. This is identical to (1), based on group export weights. For a small country of this type, it might also be reasonable to assume that foreign supply elasticities are infinite, so that (2) is useful for import prices.

The second case is a small country where the elasticity of demand tends to minus infinity while the elasticity of substitution remains finite, keeping to the assumption of differentiated products Then it can be shown that import prices remain constant in terms of foreign currency, or $P^{*k} = T_k^*$.[5] Export prices are determined by the inverse of equation (13), so that inverse import weights are appropriate rather than export weights.

In the third and most unrealistic case, cross-elasticities of demand for the country's product are assumed to be zero in all foreign markets, while own elasticities are finite and the domestic supply elasticity is zero In the relevant modification of (7) for these conditions, the export shares W_j^k in (16) should be weighted according to the elasticity of demand, that is, replaced by $\eta^k W_j^k$ The simple export weights could only be appropriate if the demand elasticities were the same in all markets

2. Effective Exchange Rates

Let us now consider an effective exchange rate, using the MERM solution (12) According to Artus and Rhomberg (1973, p 606), "the effective exchange rate change of a currency can be defined as the change that would induce the same alteration in its trade balance expressed in the numeraire currency as that brought about by a given realignment of all exchange rates" Following this definition, it is possible to derive the effective change rate E_k^* explicitly in terms of the price changes given by (12)

Since Z_j^k is the volume of exports from country j to country k, the trade balance in terms of dollars of country k is (with imports on the left-hand side and exports on the right-hand side)

$$\sum_{j\neq k} P_j Z_j^k - \sum_{j\neq k} P_k Z_k^j \qquad (18)$$

Differentiating with respect to time and using the definition $X_j^k = P_j Z_j^k$, we find

$$\sum_{j\neq k} X_j^k (P_j^* + Z_j^{*k}) = \sum_{j\neq k} X_k^j (P_k^* + Z_k^{*j}) \qquad (19)$$

Let us define $X_j = \sum_{k=1}^{n} X_j^k = \sum_{k\neq j} X_j^k + X_j^j$ Then equation (7) can be rewritten, after multiplying through by X_j and allowing for the virtual effective exchange rate change E_k^*, as

$$\sum_{j=1}^{n} X_k^j Z_k^{*j} = \sum_{j\neq k} X_k^j Z_k^{*j} + X_k^k Z_k^{*k} = X_k \beta_k (P_k^* - T_k^* - E_k^*) \qquad (20)$$

---

[5] Under these conditions, the cross-elasticities become proportional to the import shares $(\eta_{jl}^k \to \eta^k S_l^k)$, as can be seen from Armington's eq (26) As $\eta^k \to -\infty$, the relevant modification of (7) becomes $\sum_{k\neq j} W_j^k P^{*k} = \sum_{k\neq j} W_j^k T_k^*$ for all i, using (13) This will hold for all j only if $P^{*k} = T_k^*$ I am indebted to Howard Petith and Rudolf Rhomberg for pushing me to explore these three cases

Substituting from (20) into the right-hand side of (19), one finds

$$\sum_{j\neq k} X_j^k(P_j^* + Z_j^{*k}) = X_k\beta_k(P_k^* - T_k^* - E_k^*) - X_k^k Z_k^{*k} + \sum_{j\neq k} X_k^j P_k^* \tag{21}$$

or

$$\sum_{j=1}^{n} X_j^k(P_j^* + Z_j^{*k}) = X_k\beta_k(P_k^* - T_k^* - E_k^*) + \sum_{j=1}^{n} X_k^j P_k^*. \tag{22}$$

Using (8), one can rewrite (4) to include the virtual effective rate change E_k^* as

$$Z_j^{*k} = \sum_{i=1}^{n} \eta_{j/i}^k P_i^* - \eta^k T_k^* - \eta^k E_k^* \tag{23}$$

Then, if we let $X^k = \sum_{j=1}^{n} X_j^k$, (22) becomes

$$\sum_{j=1}^{n} X_j^k\left(P_j^* + \sum_{i=1}^{n} \eta_{j/i}^k P_i^*\right) - X^k\eta^k(T_k^* + E_k^*) = X_k\beta_k(P_k^* - T_k^* - E_k^*) + X_k P_k^* \tag{24}$$

Solving for E_k^* and defining $\alpha_k = X^k/X_k$, we find the MERM effective exchange rate

$$E_k^* = \frac{1 + \beta_k}{\beta_k - \alpha_k\eta^k} P_k^* - \frac{\alpha_k}{\beta_k - \alpha_k\eta^k} \sum_{j=1}^{n} s_j^k\left(P_j^* + \sum_{i=1}^{n} \eta_{j/i}^k P_i^*\right) - T_k^* \tag{25}$$

The components on the right-hand side of (25) can be computed from the MERM solution (12), giving E_k^* as the virtual effective exchange rate change by country k that would leave its trade balance unaffected after a multilateral set of changes T_j^*. Suppose, for example, that $\alpha_k = 1$, $\eta^k = -1$, and $\beta_k = 3$. Then (25) defines the effective exchange rate change as the change in the local currency price of exports $(P_k^* - T_k^*)$ less 0.25 times the percentage change in the dollar value of local demand for imported goods that would have occurred in the absence of a change in the price of the local currency. More generally, the weights of these components in the effective rate index depend on the elasticities of demand and supply for traded goods, as in (25).

It is now possible to see the difference between the multilateral index (25) and the bilateral indexes (1) and (2). Thus, while bilateral indexes may provide useful measures of the effects of floating rates on export prices and import prices of small countries, they are not useful measures of the exchange rate change by a single country that would affect its trade balance in the same way as a given set of multilateral changes in exchange rates.

There is one simple, but highly unrealistic, case in which the bilateral and multilateral indexes are equivalent Let us add to the small-country case the additional assumptions that the home country's cross-elasticities of demand are all zero ($\eta^k_{j/i} = 0$ for $i \neq j$ and $\eta^k_{j/i} = \eta^k$ for $i = j$) and that initially trade is balanced (so that $\alpha_k = 1$) If we use results (15) and (17) for the small-country case, equation (25) simplifies to

$$E_k^* = \frac{1 + \beta_k}{\beta_k - \eta^k} \sum_{j \neq k} E_{kj} T_j^* - \frac{1 + \eta^k}{\beta_k - \eta^k} \sum_{j=1}^{n} S_j^k T_j^* - T_k^*, \qquad (26)$$

a weighted average of the export-weighted and import-weighted indexes, with weights summing to unity

References

Armington, P S "A Theory of Demand for Products Distinguished by Place of Production" *IMF Staff Papers* 16 (March 1969) 159–76

Artus, J, and Rhomberg, R "A Multilateral Exchange Rate Model" *IMF Staff Papers* 20 (November 1973) 591–611

Economic Report of the President Washington Government Printing Office, 1974

Kemp, M. "The Balance of Payments and the Terms of Trade in Relation to Financial Controls" *Rev Econ Studies* 37 (January 1970) 25–31

Triangular Graphs

Joe Latham

University of Glasgow

Historical Introduction

Toward the end of the nineteenth century, extensive investigations were made of the properties of mixtures of two or more metals (alloys) based on the phase rule, which was first formulated in 1873 by Gibbs (1961, pp 55–353)

A definition of a mixture is that it possesses one or more properties which are functions only of the proportions of the components present and are independent of the total amount of the mixture For example, the hardness or melting point of an alloy is determined by the relative amounts of the components and does not depend on the amount of the alloy

Two-component mixtures can be adequately represented by a straight line drawn on a rectangular Cartesian graph The treatment of the general case involving mixtures of more than two components requires more sophistication A first step is the representation of three-component mixtures by means of triangular graphs

An early example is provided by the work of Fisher (1906), who wanted a simple means of calculating the fraction of a patient's intake of calories that originated from protein, carbohydrate, and fat In counting calories, one can regard food as a mixture of these three components

Requiring a method that could be used in a hospital ward, Fisher rightly rejected algebra and devised an ingenious technique based on triangular graphs Unfortunately, at this time, triangular coordinate graph paper was not available Fisher used square-grid coordinate paper based on an isosceles right-angled triangle, which necessarily introduces arbitrary asymmetry, so this method was not further developed

It is perhaps noteworthy that Fisher was supervised at Yale by J Willard Gibbs, which may help to explain Fisher's unusual choice of technique in dealing with this three-component-mixture problem

In the early decades of the twentieth century, the value of the work of Gibbs on the phase rule was widely recognized among metallurgists and physical chemists, leading to extensive studies of alloys containing

[*Journal of Political Economy* 1976 vol 84, no 3]
© 1976 by The University of Chicago All rights reserved

three components (Hume-Rothery, Christian, and Pearson 1952, pp. 245–53) As a result, a steady demand was created for triangular coordinate graph paper, which is now commercially available.[1]

This communication is intended to draw attention to the potential usefulness for expository and pedagogic purposes of graphs drawn on triangular coordinate paper in connection with some problems in economics For this purpose, two examples based on economic data are presented The first example shows how the composition of GNP for selected countries has changed during the period 1958–67 The second example is based on international trade from 1946–69

Such graphs cannot be used for estimation purposes, although techniques have been developed for fitting regression models to data where the explanatory variables form a mixture, such as the three-component mixtures discussed in this note (Cox 1971) A review on the general literature of mixtures is given by Cornell (1973)

Use of Triangular Graphs

From time to time, the properties of the equilateral triangle have been applied by various workers to data from outside the physical sciences Some examples, but not an exhaustive list of such applications, can be found dealing with topics such as urban reconstruction (Burns 1961, pp 83–86), the importance of the service sector in EEC countries (Lengellé 1966, pp 51–57), medical diagnosis (Aitchison, Taylor, and McGirr 1971, Aitchison 1974), and balance-of-payments adjustment (Cooper 1968*a*, pp 17–21, 1968*b*)

Nevertheless, I have found no source which attempts to demonstrate the intrinsic value of triangular graphs in economics, and it is certainly true that many economists have not seen or heard of such graphs and are unaware of their potential value

However, economics abounds in three-way classifications in which data can most simply be represented by using triangular graphs For example, the national product is often broken down into the contributions of agriculture, industry, and services A similar classification is used for the analysis of household expenditures (food, manufactured goods, services) Factors of production are often grouped as land, labor, and capital, expenditure on the national product is made up of domestic expenditure plus exports minus imports, statistics of international trade are aggregated by the United Nations into three categories, Class I (developed market economies), Class II (developing market economies), and Class III (centrally planned economies), urban building land is divided into commercial, industrial, and residential sectors The potential list could easily be extended.

---

[1] Triangular coordinate graph paper is obtainable from H W Peel Ltd , Chartwell House, Jeymer Drive, Greenford, Middlesex, United Kingdom, and from Keuffel and Esser Co , 40 East 43d Street, New York, New York 10017

Triangular graphs are somewhat analogous to scatter diagrams in providing quick screening devices which help the investigator to decide whether the plotted data have features that require further, more elaborate analysis The graphs can be helpful as pedagogic devices, particularly when showing trends in economic time series of variables that can be broken down into three component parts

Plotting Triangular Graphs

The mathematical basis of triangular graphs is that the algebraic sum of the three perpendiculars from any point to the sides of an equilateral triangle is independent of the position of the point If the point is outside the triangle, one or more of the perpendiculars will have a negative length

Consequently, any three fractions adding to unity, or three percentages totaling 100, can be represented by a single point in an equilateral trianglar graph, the lengths of the three perpendiculars being proportional to the three fractions or percentages plotted

If the three quantities being described are *A*, *B*, and *C*, then one vertex of the triangle will correspond to 100 percent of *A* and 0 percent of *B* and *C* The base *BC* corresponds to 0 percent of *A* Lines parallel to *BC* correspond to a fixed percentage of *A* which can be read from the scale printed on the graph paper The line corresponding to 110 percent of *A* (which may be needed if *B* or *C* is negative) is parallel to *BC* but outside the original triangle Similarly, the vertices *B* and *C* correspond to 100 percent of *B* and *C*, respectively

The point that represents p_a percent of *A*, p_b percent of *B*, and p_c percent of *C* is therefore the intersection of three lines The first line is parallel to *BC*, and its distance from *BC* is p_a percent of the distance from *BC* to *A* Similarly, the second line would be p_b percent of the distance from *AC* to *B* and the third line would be p_c percent of the distance from *AB* to *C*

On the standard form of triangular graph paper, each side is divided into 100 units, which is particularly convenient for expressing percentages, which may be plotted directly Time trends are shown by plotting a series of points The trends are characterized by the length of the path and the direction of movement, as well as by the position of the path in the triangle

Example 1: The Changing Structure of GNP

Figure 1 shows the change in economic structure of GNP during the period 1958–67 for 25 countries, based on a threefold classification into agriculture, industry, and services (United Nations 1968, pp 88-99)

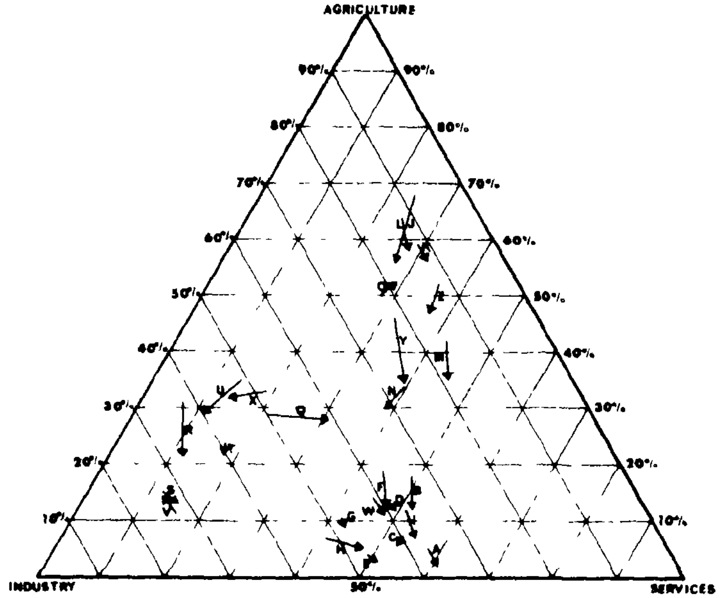


FIG. 1.—The changing structure of GNP, 1958–67. *A*, United States, *B*, Japan, *C*, Belgium, *D*, Union of South Africa, *E*, United Kingdom, *F*, Italy, *G*, France, *H*, West Germany, *I*, Norway, *J*, Nigeria, *K*, Malawi, *L*, Uganda, *M*, Kenya, *N*, Morocco, *O*, India, *Q*, Yugoslavia, *R*, Hungary, *S*, Czechoslovakia, *T*, USSR, *U*, Romania, *V*, East Germany, *W*, Australia (1958–66), *X*, Bulgaria, *Y*, Turkey, *Z*, Pakistan.

The point corresponding to 1967 is shown by the head of the arrow, whereas that for 1958 is represented by the tail of the arrow. Examination of figure 1 shows that the countries can be characterized by three features.

i) *The Region of the Graph*

It is immediately clear from the graph (but not from the corresponding tables of data) that the countries listed fall into three distinct groups: LDCs (high agriculture share), centrally planned economies (low service share), and advanced capitalist countries (low agriculture and high service share).[2] Thus, Japan is clearly grouped with the advanced capitalist countries, whereas Turkey is grouped with the LDCs.

ii) *The Length of the Arrow*

Visual examination of figure 1 suggests that the length of the "development arrow" is less for the advanced countries than for the other two groups. But some care should be exercised in detailed intercountry comparisons, as the arrow corresponding to a change from, say, 55 to 45 percent agriculture would have the same length as one corresponding to the change from 10 to 0 percent agriculture, although these two changes would clearly have different interpretations Nevertheless, a cursory examination of figure 1 indicates that the arrows of the low per capita income countries are longer than those of the advanced countries The graph thus suggests that the relation of per capita GNP to changes in the composition of GNP might be a topic worthy of further investigation, with the tentative conclusion that the developing countries are undergoing structural change more rapidly than the advanced countries

---

[2] The low share of services in the centrally planned economies is partly a matter of definition, as government services are excluded from the national product.

The length of the arrow can be conveniently computed from the coordinates (A_1, B_1, C_1) and (A_2, B_2, C_2) of the points corresponding to the head and tail of the arrow If the length of the arrow is D, it follows by simple geometry that

$$D^2 = (B_1 - B_2)^2 - (A_1 - A_2)(C_1 - C_2)$$

iii) *The Direction of the Arrow*

The trend with time of the composition of GNP is shown by the direction of the arrow, the most noticeable feature being the move away from agriculture toward industry and services As both the length and direction of the arrow can be measured, these properties may form the basis of indices of the extent and direction of structural change

If required, the direction of the arrow can be computed from the formula

$$\tan \theta = \frac{\sqrt{3}(B_1 - B_2)}{(C_1 - C_2) - (A_1 - A_2)},$$

where, as before, the two points are (A_1, B_1, C_1) and (A_2, B_2, C_2) and θ is the angle relative to the base of the triangle opposite the apex numbered B (i e , the direction from apex A to apex C is zero degrees)

It is immediately apparent from figure 1 that the arrows corresponding to the developing countries are moving toward the region of the advanced capitalist countries In this limited sense, the graph gives one aspect of the development process

Example 2: The Changing Structure of World Trade

The second example shows the application of triangular graphs to the analysis of trends in world trade (United Nations 1969, p 12) The data plotted are the source of imports entering international trade over the period 1948 69 The classification used is that of the United Nations, namely, Class I = developed market economies Class II = developing market economies, and Class III = centrally planned economies

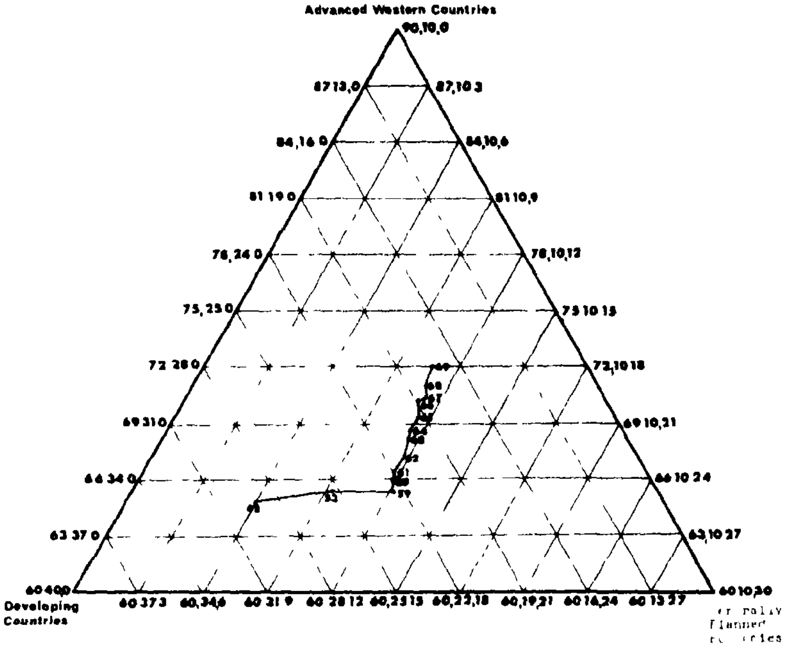


Fig 2 —The changing structure of world trade, 1946–69 This graph is plotted on an expanded scale The set of three numbers shown along the sides of the triangle gives the percentages of the world's imports into Class I, Class II, and Class III countries, respectively, at the corresponding points on the sides of the triangle

In figure 2, the triangular scale has been expanded by a factor of 3⅓, so that the three sides of the triangle are represented by the lines Class I, imports = 60 percent, Class II, imports = 10 percent, and Class III, imports = 0 percent The full length of the side of the triangle corresponds to a change of 30 percent in one of the values plotted (The sets of three numbers marked along the sides of the triangle give percentages of imports for Class I, Class II, and Class III, respectively)

The graph shows the general movement of world trade away from the developing countries, with a sharp discontinuity in 1959 when trade moved sharply toward the developed market economies

The discontinuity could be an artifact if it were caused by some change in the method of reporting over the period In this connection, it may be significant that the change in the reporting interval coincides with the discontinuity. Another possible explanation is the rapid development of intra-European trade during the period which followed the formation of the European Common Market Whatever the explanation, the graphical presentation immediately focuses attention on the discontinuity, and so reveals a problem that might otherwise remain hidden in the tables of data It is in such preliminary examinations of data that triangular graphs may have their greatest value

Other Uses of Triangular Graphs

These graphs may be used wherever a property can be broken down into three component parts which add up to a constant total These parts may be individually positive or negative, for example, the relation

$$\text{domestic product} = \text{domestic expenditure} + \text{exports} - \text{imports}$$

can be plotted if domestic expenditure, exports, and imports are expressed as percentages of the domestic product The corresponding point will lie outside the original triangle opposite the apex corresponding to imports The base of the triangle would be the line

$$\text{imports} = 0$$

The plotted points would be in the region of the graph farther from the imports apex than this base line

Triangular graphs can also represent three quantities whose product is constant by the simple expedient of using the logarithms of the properties as the basic variable, since

$$\log (XYZ) = \log (X) + \log (Y) + \log (Z)$$

An example where this can be useful occurs if corresponding indices of price (P) and of quantity (Q) are available for a number of years The value (V) in real terms is given by

$$V = P \times Q \times F,$$

where F is a deflating factor

One curiosity which is perhaps noteworthy is that, since the equation of the surface of a sphere is

$$x^2 + y^2 + z^2 = r^2,$$

it is possible to represent points on a spherical surface on triangular graph paper As a result, maps of the earth's surface can be drawn in which both latitude and longitude are exact straight lines Such a triangular mapping would represent one octant of the earth's surface (say, from the north pole to points of longitude 0° and 90° E on the equator)

References

Aitchison, John "Hippocratic Inference" *Bull Inst Math and Its Appl.* 10 (January/February 1974) 48–53

Aitchison, John, Taylor, Thomas R ; and McGirr, Edward M "Doctors as Decision Makers A Computer Assisted Study of Diagnosis as a Cognitive Skill" *British Medical J* 3 (July 1971) 35–40

Burns, Leyland S *Investment Time Scheduling in Urban Reconstruction* 's Gravenhage Drukerij Pasmans, 1961

Cooper, Richard N *The Economics of Interdependence* New York McGraw-Hill, 1968 (*a*)

——— "Relevance of International Liquidity to Developed Countries" *A E R* 58 (May 1968) 627 (*b*)

Cornell, John A "Experiments with Mixtures A Review" *Technometrics* 15, no 3 (1973) 437–55

Cox, D R "A Note on Polynomial Response Functions for Mixtures" *Biometrika* 58, no 1 (1971) 155–59

Fisher, Irving "A New Method for Indicating Food Values" *American J Physiology* 15 (1906)· 417–32.

Gibbs, J Willard *The Scientific Papers of J Willard Gibbs* Vol 1, *On the Equilibrium of Heterogeneous Substances* New York Dover, 1961

Hume-Rothery, William, Christian, J S , and Pearson, W B "The Study of Ternary Systems" In *Metallurgical Equilibrium Diagrams* London Inst Physics, 1952

Lengellé, Maurice *The Growing Importance of the Service Sector in Member Countries* Paris Org Econ Cooperation and Development, 1966

United Nations *Yearbook of National Account Statistics* Vol 2 New York United Nations, 1968

——— *Yearbook of International Trade Statistics* New York United Nations, 1969

The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics

Thomas J. Sargent

University of Minnesota

The usual proof that Friedman's simple *k*-percent growth rule for the money supply is suboptimal comes from mechanically manipulating a reduced-form equation. Those manipulations, in general, show that pursuing a rule with feedback from current economic conditions to the money supply is better than following Friedman's advice. To be valid, the proof requires that, as written in one particular way, the reduced-form equation will remain unaltered when the monetary authority departs from the old "rule" used during the estimation period and follows a new one. Here I point out that there are always alternative ways of writing the reduced form, one being observationally equivalent with the other, so that each is equally valid in the estimation period. If one assumes that the first form is invariant when the policy rule is changed, the proof of the superiority of rules with feedback over Friedman's rule goes through. But if one assumes that it is the reduced form as written in the second way that remains unchanged, the proof that Friedman is wrong does not obtain—instead, the implication is that Friedman's rule does as well as any other deterministic feedback rule and better than a stochastic rule. Therefore, estimates of reduced forms alone will not permit one to settle the difference between Friedman and advocates of rules with feedback. Given any set of reduced-form estimates, there is an invariance assumption that will permit a member of either camp to make his point. In effect, then, this paper poses the question: Does the view that Friedman's *k*-percent feedback rule is as good as any other deterministic feedback rule place any restrictions on reduced forms? The answer is no. This is distressing since, for a given sampling interval and estimation period, the reduced-form estimates summarize everything that the data can ever tell

John Geweke, Christopher Sims, Gary Skoog, and Neil Wallace contributed valuable comments on an earlier draft of this paper. The Federal Reserve Bank of Minneapolis financed the work. The views expressed herein are solely my own and do not necessarily represent the views of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

[*Journal of Political Economy*, 1976, vol. 84, no. 3]
© 1976 by The University of Chicago. All rights reserved.

us To rule on the policy issue thus requires bringing to bear theoretical considerations or doing empirical work of a kind considerably more subtle than that directed solely at estimating reduced forms. In effect, it is necessary to get some evidence on what sort of invariance assumption is the most realistic one to impose How one does that is a delicate, though not entirely intractable, task, the discussion of which is outside the scope of this paper

This paper is in the nature of a footnote to Lucas's (1973*a*) important critique of econometric policy evaluation Lucas emphasized the critical invariance assumption behind the usual argument for rules with feedback and showed how that invariance assumption fails to hold in models with rational expectations As extreme examples of how wrong the standard invariance assumption could be, Lucas (1972) and Sargent and Wallace (1975) have constructed particular structural models in which one deterministic rule is as good as another, so that the standard proof of the suboptimality of Friedman's rule fails spectacularly in those examples Those examples were dependent on particular structural setups The point of this note is that, for any estimated reduced form, there is an invariance assumption which if imposed delivers the conclusion that one deterministic rule is equivalent with any other In effect, then, this note displays some mechanical equivalencies that force one to stumble upon Lucas's observations about the limits of the usual applications of optimal-control theory to macroeconometric models

A casual reader of Marschak's classic paper (1953) would perhaps regard my major contention—that reduced forms alone cannot settle the policy-rules controversy—as being obvious However, applications of optimal control to macroeconometric models do purport to extract implications about the optimal feedback rule solely from estimated reduced forms

For simplicity, I deal with a bivariate model I assume that during the estimation period, two variables, y_t and m_t, were described as a realization from a stationary, indeterministic stochastic process The variable y_t measures some "goal" variable like unemployment or GNP The variable m_t represents a potential policy instrument I assume that during the estimation period m_t was exogenous with respect to y, and that m_t caused y_t, in Granger's sense (1969) This means that the (y, m) process can be represented in the particular (Wold) moving average form[1]

$$y_t = \alpha(L)\varepsilon_t + \beta(L)\eta_t \tag{1a}$$

$$m_t = \gamma(L)\varepsilon_t, \tag{1b}$$

---

[1] The paper by Sims (1972), especially his appendix, provides a useful summary of the statistical theory used here

where

$$\alpha(L) = \sum_{j=0}^{\infty} \alpha_j L^j, \qquad \beta(L) = \sum_{j=0}^{\infty} \beta_j L^j,$$

and

$$\gamma(L) = \sum_{j=0}^{\infty} \gamma_j L^j,$$

L is the lag operator ($L^n x_t \equiv x_{t-n}$), and η_t and ε_t are mutually uncorrelated and serially uncorrelated random variables with means of zero and finite variances σ_η^2 and σ_ε^2, respectively (I have omitted constants and any deterministic terms from representation [1], which can be included without affecting the argument.)[2]

The assumption that the Wold representation has the triangular form of (1)—that is, the assumption that m is exogenous and "causes" y—means that under one pretty general additional condition, a final-form regression of y on m can be consistently estimated by least squares. In particular, suppose that $\gamma(L)$ is invertible so that m_t has the autoregressive representation $\gamma^{-1}(L)m_t = \varepsilon_t$, where $\gamma^{-1}(L)$ is a one-sided polynomial in the lag operator L. Substituting the above equation into (1a) gives

$$y_t = \alpha(L)\gamma^{-1}(L)m_t + \beta(L)\eta_t \tag{2}$$

$$y_t = h(L)m_t + \beta(L)\eta_t,$$

where $h(L) = \alpha(L)\gamma^{-1}(L)$ is a one-sided polynomial in the lag operator L. Equation (2) is a "final form" for y in terms of m and is consistently estimated by least squares, since the η_t process is orthogonal to the m_t process. Assuming that $\beta(L)$ has a one-sided inverse, the "reduced form" for y_t can be obtained as[3] $\beta(L)^{-1}y_t = \beta(L)^{-1}h(L)m_t + \eta_t$, or

$$y_t = \sum_{i=0}^{\infty} a_i m_{t-i} + \sum_{i=0}^{\infty} b_i y_{t-i-1} + \eta_t, \tag{3}$$

where

$$a(L) = \sum_{i=0}^{\infty} a_i L^i = \beta(L)^{-1}h(L)$$

and

$$[1 - Lb(L)] = \left(1 - \sum_{i=0}^{\infty} b_i L^{i+1}\right) = \beta(L)^{-1}$$

---

[2] I will also assume that ε and η are mutually and serially independent, which facilitates the computations below but is not essential. Abandoning that assumption would require replacing mathematical expectations with linear least-squares forecasts at several points below. With that replacement, my argument would go through.

[3] I am assuming that $\beta(L)$ and σ_η^2 are so normalized that $\beta_0 = 1$, which implies that the zero-order coefficient of $\beta(L)^{-1}$ is also unity.

The reduced form (3) expresses y in terms of current and lagged m's and lagged y's with a disturbance that is serially uncorrelated and orthogonal to the variables on the right-hand side of the equation. Thus the reduced form (3) can be consistently estimated by least squares.

Write (3) as

$$y_t = a(L)m_t + b(L)y_{t-1} + \eta_t. \tag{4}$$

It is now easy to illustrate the elements of the usual argument that it is optimal for the policy authority to set m via a rule with feedback from lagged y's to current m. Suppose that the authority's goal is to set m in order to minimize the variance of y. Suppose that the parameters of the reduced form (4) will remain unaltered when the authority abandons (1b), which described policy during the estimation period, and implements a new feedback rule. (Suppose also that $a(L)$ has a one-sided inverse under convolution, which is not really necessary but rules out unseemly "instrument instability" problems.) Then it is straightforward to show that the authority would minimize the variance of y by using the feedback rule

$$a(L)m_t = -b(L)y_{t-1}, \tag{5}$$

which is a rule for setting m_t as a function of lagged m's and lagged y's. In general, some of the b_j's are not zero, so that it is optimal for the authority to incorporate feedback from lagged y's to m, presumably to "lean against the wind." So a rule without feedback is suboptimal.

As Lucas (1973*b*) has emphasized, a key assumption in the above argument for rules with feedback is that the parameters of the reduced form (4) remain unchanged when the authority abandons the rule used during the estimation period and uses a new one. Lucas argued that if the reduced form incorporates the influence of people's expectations, and if expectations are formed in a well-informed or "rational" way, that assumption is not appropriate. Here I point out that there always seems to be an interpretation of the reduced form (4) which completely vitiates the preceding demonstration of the superiority of rules with feedback.

I begin by noting that from (1b) $\gamma_0\varepsilon_t = m_t - E_{t-1}m_t$, where $E_{t-1}m_t$ is the mathematical expectation of m_t conditioned on past m's and past y's. In this case, since m is exogenous in the estimation period, lagged y's don't help explain m once lagged m's are taken into account, so that in the estimation period $E_{t-1}m_t \equiv E(m_t \mid m_{t-1}, m_{t-2}, \ldots, y_{t-1}, y_{t-2}, \ldots) = E(m_t \mid m_{t-1}, m_{t-2}, \ldots)$. The random variable $\gamma_0\varepsilon_t$ is the "innovation" of m_t—the part that can't be predicted on the basis of past m's (or past y's). Now substituting $\varepsilon_t = (1/\gamma_0)(m_t - E_{t-1}m_t)$ into (1a) and rearranging gives $\beta(L)^{-1}y_t = (1/\gamma_0)\beta(L)^{-1}\alpha(L)(m_t - E_{t-1}m_t) + \eta_t$, or

$$y_t = c(L)(m_t - E_{t-1}m_t) + b(L)y_{t-1} + \eta_t, \tag{6}$$

where

$$\frac{1}{\gamma_0}\beta(L)^{-1}\alpha(L) = c(L) = \sum_{i=0}^{\infty} c_i L^i,$$

and where $b(L)$ is as defined under (3).[4] Equation (6) is an alternative version of the reduced form that is equivalent with the version (3) or (4) from the point of view of representing things during the estimation period. Notice that (4) and (6) have identical residuals and so fit equally well.

If we assume that the reduced form (6) remains unaltered when the authority abandons (1b) and adopts a deterministic feedback rule giving m as an exact function of past m's and past y's, a strong sort of "neutrality" result emerges. For under any deterministic feedback rule, say one of the form

$$m_t = s_1 m_{t-1} + s_2 m_{t-2} + \cdots + r_1 y_{t-1} + r_2 y_{t-2} + \cdots, \tag{7}$$

it is true that

$$E_{t-1} m_t = m_t, \tag{8}$$

identically in t. Substituting (8) into (6) gives $y_t = b(L) y_{t-1} + \eta_t$, which is an autoregressive representation for y that holds regardless of the values of the particular parameters $s_1, s_2, \ldots, r_1, r_2, \ldots$ of the feedback rule (7) selected by the authority. The assumption that it is the reduced-form representation (6) that remains unchanged as the policy rule is altered leads to the conclusion that one deterministic feedback rule is as good as any other. There is thus no reason to expect that the authority can do better than it can by implementing the x-percent growth rule recommended by Friedman.

The preceding "neutrality" demonstration rests on the arbitrary assumption that it is the reduced-form representation (6) that remains unaltered when the authority institutes a new policy rule outside the estimation period. Of course, the earlier demonstration of the superiority of a rule with feedback depended on the equally arbitrary but different assumption that it was the reduced-form representation (4) that remained unchanged from the estimation period even once the new policy rule was instituted. From the viewpoint of extracting policy implications, assuming invariance for the reduced-form representation (4) or (6) thus gives drastically different implications. Yet, from the point of view of representing the reduced form during the estimation period, (4) and (6) are exactly equivalent. This is what leads me to the conclusion that the empirical evidence from a single estimation period alone, which can be completely summarized by (4) or (6) and an autoregression for *m*, can never settle things between advocates of rules with feedback and advocates of rules without feedback

---

[4] Notice that $c(L)(m_t - E_{t-1}m_t) = c_0(m_t - E_{t-1}m_t) + c_1(m_{t-1} - E_{t-2}m_{t-1}) + c_2(m_{t-2} - E_{t-3}m_{t-2}) + \cdots$.

Perhaps this could be dismissed as a mere curiosity if macroeconomists agreed that as between (4) and (6) one of these ways of writing the reduced form is much more likely to remain unchanged when policy changes The problem is that there is no such agreement. While in the past most macroeconomists have regarded (4) as invariant under changes in the policy rule, that assumption depends critically on the assumption that peoples' expectations are formed using fixed-weight, autoregressive schemes that in general are not "rational" Lucas (1973*b*) has argued forcefully against that assumption, but it remains true that the assumption still underlies most macroeconometric policy evaluation and is an essential element of most applications of control theory to macroeconometric models On the other hand, reduced-form representations resembling (6) are supposed to be invariant under changes in the policy rule according to some structural macroeconomic models incorporating rational expectations and Lucas's (1973*a*) formulation of the aggregate supply function [5] In such models, current and maybe lagged innovations in the money supply are what agents respond to

The upshot is that the invariance of neither (4) nor (6) to changes in the policy rule would now command a concensus among macroeconomists The current state of macroeconomic theory seems to me to be very far from supplying a reliable basis for ruling out one of these invariance assumptions in favor of the other For that reason, I believe that the observational equivalence of (4) and (6) provides some cause to be circumspect about economists' ability to be sure that rules with feedback clearly dominate rules without feedback (or vice versa)

The reader may wonder whether my assumptions that *m* is exogenous with respect to *y* and that *m* causes *y* in effect rig things in the preceding argument They don't Those assumptions were made to guarantee that a *y*-on-*m* reduced form was identifiable and estimable The estimability of such a reduced form is a sine qua non for the usual argument that rules with feedback dominate rules without feedback One alternative set of assumptions would have been that *y* and *m* caused each other, so that there was mutual feedback between *y* and *m* Only under special circumstances, an instance of which is analyzed in Appendix *B*, is a *y*-on-*m* reduced form identified in a system with mutual feedback As Appendix B illustrates, for systems with mutual feedback that are identifiable through a priori restrictions, there obtains the same observational equivalence as analyzed in the text Another alternative assumption would have been that *y* was exogenous with respect to *m* in the estimation period. I have elsewhere advanced the notion that the hypothesis of exogeneity of certain goal variables y (e g , unemployment) with respect to certain policy instruments m (e g , the money supply) is a naive, model-free way of stating the natural-rate hypothesis [6] Exogeneity of y with respect to m can readily be shown to be compatible with the notion that one deterministic rule is as good as any other on a certain invariance assumption On the other hand, it is straightforward to show that an alternative invariance assumption could be imposed that would imply that Friedman's rule is suboptimal This is shown in Appendix A

---

[5] For example, see Sargent and Wallace (1975)

The present argument only shows how reduced forms estimated for a given sampling interval (i e , quarterly or monthly) over a given estimation period cannot settle the policy-rules controversy. That does not mean that there is no way that empirical evidence can be brought to bear on the question Presumbly, by estimating reduced forms for various subperiods or countries across which policy rules differed systematically, light can be shed on what way of writing the reduced form remains invariant [7] Alternatively, by studying data more and less finely aggregated over time, different implications of our two invariance assumptions might be extracted and tested Both of these paths involve considerable subtleties Very little satisfactory evidence has yet been assembled along either path

Appendix A

How "Classical" Models Can Be Interpreted in "Keynesian" Ways

This Appendix illustrates how, for classical models in which real variables are econometrically exogenous with respect to policy variables, there is a way of writing the reduced form which, if invariant under rules changes, implies that rules with feedback are optimal Therefore, evidence that real variables are econometrically exogenous with respect to policy variables, which I have argued (Sargent 1976) is a strong, model-free way of stating the natural-rate hypothesis, has "classical" policy implications on one kind of invariance assumption but does not on another The argument in this Appendix is thus the other half of the observational equivalence dilemma, since here I start with a model originally thought to be very "neutral" or classical and produce an invariance assumption that rationalizes rules with feedback

Consider a "structural" model of the form

$$\lambda(L)y_t = \gamma(m_t - E_{t-1}m_t) + u_t, \gamma > 0 \qquad (A1)$$

$$m_t = d(L)\varepsilon_t. \qquad (A2)$$

---

[6] See Sargent (1976)

[7] Lucas's (1973a) international comparisons provide an excellent example of this approach

where

$$\lambda(L) = \sum_{i=0}^{\infty} \lambda_i L^i,\ d(L) = \sum_{i=0}^{\infty} d_i L^i,$$

$\lambda(L)$ and $d(L)$ both have one-sided inverses under convolution, and u_t and ε_t are mutually and serially independent random variables with means of zero and finite variances The model (A1)-(A2) is a two-variable example of the "classical" model described by Sargent (1976) Here innovations in m (money) produce sympathetic movements in y. In this model, y is exogenous with respect to m, and m is exogenous with respect to y Though y and m are correlated, neither one helps predict the other To see that m is exogenous with respect to y, notice that (A1) and (A2) can be rearranged in the triangular Wold representation,

$$y_t = \lambda^{-1}(L)\gamma d_0 \varepsilon_t + \lambda^{-1}(L)u_t \tag{A3}$$

$$m_t = d(L)\varepsilon_t, \tag{A4}$$

the existence of which shows that m is exogenous with respect to y by virtue of Sims's theorem 1 To see that y is exogenous with respect to m, observe directly from (A1) that

$$E_{t-1} y_t = -\sum_{i=1}^{\infty} \lambda_i y_{t-i},$$

which shows that m does not help predict y once lagged y's are accounted for Alternatively, rewrite (A1)-(A2) as

$$y_t = \lambda^{-1}(L)w_t \tag{A5}$$

$$m_t = d(L)\phi w_t + d(L)\xi_t, \tag{A6}$$

where $w_t \equiv \gamma d_0 \varepsilon_t + u_t$, and ξ_t and ϕ obey $\varepsilon_t = \phi w_t + \xi_t$, $E(\xi_t | w_t) = 0$ (ϕ is the regression coefficient of ε_t against w_t, ξ_t being the residual) Since ξ_t is orthogonal to w_t and since both are serially uncorrelated by construction, it follows that (A5)-(A6) is a Wold representation for the $(y_t - m_t)$ process The existence of such a triangular Wold representation establishes that y is exogenous with respect to m, again by virtue of Sims's theorem 1

Inverting (A4) and substituting it into (A3) delivers $\lambda(L)y_t = \gamma d_0 d(L)^{-1} m_t + u_t$, which is a y – on – m reduced form which is consistently estimated by least squares, since u_t is orthogonal to lagged y's and current and lagged m's This is exactly the form of reduced form manipulated and assumed invariant under alternative policy rules in the proof that Friedman's rule is suboptimal.

If the reduced form (A1) is invariant under rules changes, then one deterministic rule is as good as any other. Thus, there are alternative ways of deriving policy implications from empirical evidence generated by a "classical" model like (A1)-(A2) Depending on what sort of invariance assumption is imposed, drastically different inferences follow about the implications of different feedback rules

Appendix B

Observational Equivalence in the Presence of Feedback from y to m

Suppose that during the estimation period the $y - m$ process possessed the vector autoregressive representation

$$y_t = a(L)y_{t-1} + b(L)m_{t-1} + \varepsilon_t \tag{B1}$$

$$m_t = c(L)y_{t-1} + d(L)m_{t-1} + u_t, \tag{B2}$$

where

$$a(L) = \sum_{j=0}^{\infty} a_j L^j,$$

etc., and ε_t and u_t are serially independent random variables with means zero and finite variance. In general, $E(\varepsilon_t u_t) \neq 0$, although it is easy to prove that $E(\varepsilon_t u_s) = 0$ for $t \neq s$. The last equality follows because, for example, u is the residual in a projection of m on lagged m's and y's and so is orthogonal to them. Since lagged u's and ε's are linear combinations of lagged y's and m's, it follows that current u is orthogonal to lagged u's and ε's. Here ε_t and u_t are the one-step-ahead prediction errors for y_t and m_t, respectively, both predictions being conditional on lagged m's and lagged y's. Consistent with usual usage, by a $y-$on$-m$ reduced form I mean a regression of y on past y's and current and past m's. Analogously, by a $m-$on$-y$ reduced form I mean a regression of m on past m's and current and past y's. For a system with probability distribution characterized by (B1)-(B2), the implied pair of $y-$on$-m$, $m-$on$-y$ reduced forms is identifiable only if sufficient a priori information is imposed on the covariance between the disturbances in the two reduced forms and on the contemporaneous coefficient in either the $y-$on$-m$ reduced form or the $m-$on$-y$ reduced form. Here I will impose the restrictions that there is no contemporaneous feedback from y to m and that the reduced-form disturbances are contemporaneously orthogonal. These restrictions serve to identify the y-m feedback structure. In particular, (B2) gives the "feedback rule" that governs m in the sample period. To find the y-on-m reduced form in terms of the parameters of (B1)-(B2), first project ε_t on u_t to get the decomposition $\varepsilon_t = \rho u_t + \xi_t$, $E(\xi_t \; u_t) = 0$, $\rho = Eu_t\varepsilon_t/Eu_t^2$. Then subtract ρm_t from (B1) and rearrange to obtain

$$y_t = [a(L) - \rho c(L)]y_{t-1} + \rho m_t + [b(L) - \rho d(L)]m_{t-1} + \xi_t \tag{B3}$$

Equation (B3) is a y-on-m reduced form with a disturbance ξ_t that is orthogonal to the regressors and also serially uncorrelated. Hence, the parameters of (B3) will be consistently estimated by least squares. Notice that (B2) is also consistently estimated by least squares.

Write (B3) more compactly as

$$y_t = h(L)y_{t-1} + g(L)m_t + \xi_t, \tag{B3$'$}$$

where $h(L) = a(L) - \rho c(L)$, $g(L) = \{\rho + L[b(L) - \rho d(L)]\}$.

Then solve (B2) for m_t in terms of lagged y's and current and lagged u's. $m_t = [1 - Ld(L)]^{-1}c(L)y_{t-1} + [1 - Ld(L)]^{-1}u_t$. Substituting this for m_t in (B3′) gives $y_t = \{h(L) + g(L)[1 - Ld(L)]^{-1}c(L)\}y_{t-1} + g(L)[1 - Ld(L)]^{-1}u_t + \xi_t$ or

$$y_t = i(L)y_{t-1} + j(L)(m_t - E_{t-1}m_t) + \xi_t, \tag{B4}$$

where $i(L) = \{h(L) + g(L)[1 - dL(L)]^{-1}c(L)\}$, $j(L) = g(L)[1 - Ld(L)]^{-1}$, and ξ_t is orthogonal to lagged y's and current and lagged u's (i.e., $[m_t - E_{t-1}m_t]$'s). So (B4) is consistently estimated by least squares.

Now (B3′) and (B4) have identical residuals and thus are observationally equivalent. The argument in the text thus goes through for this system.

References

Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods." *Econometrica* 37 (July 1969): 424–38.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24.

———. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34. (*a*)

———. "Econometric Policy Evaluation A Critique." Working Paper, Carnegie-Mellon Univ., 1973. (*b*)

Marschak, Jacob. "Economic Measurements for Policy and Prediction." In *Studies in Econometric Method*, edited by William C. Hood and T. C. Koopmans. Cowles Foundation Monograph no. 14. New Haven, Conn.: Yale Univ. Press, 1953

Sargent, Thomas J. "A Classical Macroeconometric Model for the United States." *J.P.E* 84, no 2 (April 1976) 207–37

Sargent, Thomas J., and Wallace, Neil. "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* 83, no. 2 (April 1975) 241–54.

Sims, Christopher A. "Money, Income, and Causality." *A E.R* 62 (September 1972) 540–52.

Measuring Gains and Losses

George W McKenzie

University of Southampton

In the literature of applied welfare economics, two related questions have attracted widespread interest (1) Under what conditions is it possible to construct a monetary measure of gains or losses accruing to *individual* consumers when there is only incomplete knowledge of their preferences? (2) Under what conditions is it possible to consistently *aggregate* such welfare indicators for a group of consumers so as to make meaningful statements about the effects of alternative social states on that group?

Recently McKenzie and Pearce (1976) proposed a new procedure for ranking alternative projects and policies which does not require any a priori restrictions to be placed on consumer preferences It is based on information which is, in principle, observable And it can be shown to be more accurate than either quantity variations or approximate consumer surplus measures

The purpose of this paper is to extend this approach in a number of directions First, I show that our indicator is equal to the equivalent variation Then conditions are developed for aggregating the measure across a heterogeneous group of individuals and for constructing an aggregate equivalent variation

1. Outline of the New Approach

Let us represent the preferences of the rth consumer by the indirect function

$$U_r = U_r[\phi_r(p_1, p_2, \ldots, p_n, Y_r)], \tag{1}$$

where U_r is some monotonic transformation of the function ϕ_r, Y_r is the rth individual's income, and p_i is the price of the ith commodity A change in satisfaction can then be represented by the Taylor series

$$\begin{aligned}\Delta U_r = {} & \sum_i \frac{\partial U_r}{\partial p_i}\Delta p_i + \frac{\partial U_r}{\partial Y_r}\Delta Y_r + \frac{1}{2}\sum_i\sum_j \frac{\partial^2 U_r}{\partial p_i \partial p_j}\Delta p_i \Delta p_j \\ & + \sum_i \frac{\partial^2 U_r}{\partial p_i \partial Y_r}\Delta p_i \Delta Y_r + \frac{1}{2}\frac{\partial^2 U_r}{\partial Y_r^2}(\Delta Y_r)^2 + \cdots + R_n,\end{aligned} \tag{2}$$

[*Journal of Political Economy*, 1976, vol 84 no 3]
© 1976 by The University of Chicago All rights reserved

where R_n represents the remainder associated with an nth-order approximation.

It is well known from the conditions for maximizing (1) subject to a budget constraint that the various derivatives in (2) can be rewritten as functions of quantities, prices, income effects, their higher-order derivatives with respect to prices and income, and the marginal utility of money, λ_r All of these are, of course, evaluated at the initial values of Y_r and the p_i

Now the key to the new approach is as follows As we have shown elsewhere, it is possible, without any loss of generality, to choose a monotonic transformation U_r of ϕ_r such that the initial values of its derivatives with respect to income $\partial^i\lambda/\partial Y_r^i$ ($i = 1, \quad , n$) all equal zero. This then enables the remaining derivatives of λ to be expressed in terms of price and income effects, and hence for all traces of the marginal utility of money to be eliminated from (2)

The expressions for the initial values of the various first- and second-order derivatives of the utility function can then be written as

$$\frac{\partial U_r}{\partial p_i} = -X_{ir}, \tag{3}$$

$$\frac{\partial U_r}{\partial Y_r} = 1, \tag{4}$$

$$\frac{\partial^2 U_r}{\partial Y_r^2} = 0, \tag{5}$$

$$\frac{\partial^2 U_r}{\partial p_j \partial Y_r} = -\frac{\partial X_{jr}}{\partial Y_r} = \frac{\partial \lambda_r}{\partial p_j}, \tag{6}$$

$$\frac{\partial^2 U_r}{\partial p_i \partial p_j} = -\frac{\partial X_{ir}}{\partial p_j} - X_{ir}\frac{\partial \lambda_r}{\partial p_j} = -\frac{\partial X_{ir}}{\partial p_j} + X_{ir}\frac{\partial X_{jr}}{\partial Y_r}, \tag{7}$$

where X_{ir} represents the amount of commodity i consumed by individual r Then, by substituting (3)–(7) into (2), we obtain

$$\begin{aligned}\Delta U_r = &-\sum X_{ir}\Delta p_i - \frac{1}{2}\sum_i \sum_j \left(\frac{\partial X_{ir}}{\partial p_j} - X_{ir}\frac{\partial X_{jr}}{\partial Y_r}\right)\Delta p_j \Delta p_i \\ &+ \Delta Y_r - \sum \frac{\partial X_{ir}}{\partial Y_r}\Delta p_i \Delta Y_r + \quad + R_n\end{aligned} \tag{8}$$

Carried to an infinite number of terms, this procedure enables the construction of an exact cardinal representation of an ordinal consumer preference function Obviously, such a calculation is not practicable However, the change in the cardinal indicator can be approximated to any desired degree of accuracy by adding higher-order terms to the second-order approximation contained in (8).[1] This procedure offers scope for greater accuracy than either quantity variation or consumer surplus indicators, at the same time requiring, as they do, knowledge of consumer demand functions. Now let us examine how this measure can be interpreted.

2. The Equivalent Variation

Suppose that prices and income vary in such a way that consumer satisfaction is altered by some amount, F, which has been calculated from equation (8). This same change in satisfaction can also be achieved if prices are held constant but income is changed by a sufficient amount, say ΔY^*. Thus, the change in individual r's satisfaction can also be represented by

$$\begin{aligned}\Delta U_r = F_r &= \frac{\partial U_r}{\partial Y_r}\Delta Y_r^* + \frac{1}{2}\frac{\partial^2 U_r}{\partial Y_r^2}(\Delta Y_r^*)^2 + \frac{1}{6}\frac{\partial^3 U_r}{\partial Y_r^3}(\Delta Y_r^*)^3 + \cdots + R_n^* \\ &= \lambda\Delta Y_r^* + \frac{1}{2}\frac{\partial\lambda}{\partial Y_r}(\Delta Y_r^*)^2 + \frac{1}{6}\frac{\partial^2\lambda}{\partial Y_r^2}(\Delta Y_r^*)^3 + \cdots + R_n^*,\end{aligned} \tag{9}$$

where R_n^* is the remainder association with the nth-order approximation. However, the approach described in Section 1 involves a monotonic transformation such that the initial value of λ equals 1 and the initial values of all $\partial^i\lambda/\partial Y^i$ equal 0. Thus, the change in satisfaction (8) equals ΔY_r^*, the equivalent variation. In other words, we have created a cardinal representation of an ordinal preference function where that cardinal representation can be expressed in terms of dollars or other appropriate units of currency.

3. Community Behavior Functions

The previous discussion has been carried out solely in terms of the case of a single individual. From a practical viewpoint, sufficient data are obviously not available to undertake calculations on such a basis. Therefore, it is necessary to seek conditions under which aggregative information about a particular group of individuals can be meaningfully interpreted. In the following discussion, I shall follow the dichotomy adopted by Pearce (1964). First, I examine the purely technical problem of determining those conditions under which we can construct an approximation, analogous to (8), on the basis of a community behavior function. I then turn to the conditions under which policy implications can be meaningfully drawn from such an aggregative measure.

---

[1] The magnitude of the remainder associated with any order approximation depends, of course, on the magnitude of variable changes and the size of the parameters of the consumer demand functions.

As is well known, it is in general not possible to construct a community behavior function possessing all the properties of individual preference functions As Pearce has shown (1964, p 188), provided the latter exist and the distribution of nominal money income remains unchanged, a necessary and sufficient condition for such a community behavior function to exist is that

$$\sum_r \left(X_{jr}\frac{\partial X_{ir}}{\partial Y_r} - X_{ir}\frac{\partial X_{jr}}{\partial Y_r}\right) = \left(X_j\frac{\partial X_i}{\partial Y} - X_i\frac{\partial X_j}{\partial Y}\right) = 0, \qquad (10)$$

where

$$\sum_r X_{ir} = X_i,$$

$$\sum Y_r = Y,$$

and

$$\frac{\partial X_i}{\partial Y} = \frac{\sum_r Y_r(\partial X_{ir}/\partial Y_r)}{Y}$$

If this is the case, the community behavior function can be maximized subject to an aggregate budget constraint to obtain first-order conditions analogous to (3) and (4) Higher-order derivatives can be calculated and a monotonic transformation chosen such that all traces of the community version of the marginal utility of money disappear Movements from one community behavior surface to another can then be represented by an indicator analogous to (8) where the price and income effects are now those associated with aggregate demand functions

It is now appropriate to examine the welfare significance of this procedure If the preferences of all individuals in the group under study are identical and if no redistribution of income is involved, then there is no problem An increase in the aggregate measure indicates that everyone has gained However, if preferences are not identical, there are significant difficulties Suppose that the proposed indicator reveals that there has been a move to a higher social behavior surface In the process of the move, some individuals may have gained whereas the position of others may have deteriorated However, provided that (10) holds, incomes could be redistributed in such a way as to make everyone better off An improvement in the aggregate indicator is thus a necessary but not a sufficient condition for an improvement in social welfare The income redistribution would also have to be carried out

4. The Aggregate Equivalent Variation

In Section 2, the welfare indicator for an individual is interpreted as a money measure, the equivalent variation If we wish to do the same for the aggregative version, it is necessary to impose the "adding-up" condition that the sum of all individual equivalent variations equal the value of the measure obtained on the basis of information about the aggregate demand functions

Returning to equation (8), we determine that aggregation of the second-order approximation of the equivalent variation requires that

$$\sum_r X_{ir} \frac{\partial X_{jr}}{\partial Y_r} \Delta p_j \Delta p_i = X_i \frac{\partial X_j}{\partial Y} \Delta p_j \Delta p_i \tag{11}$$

But for this to be the case,

$$\sum_r X_{ir} \frac{\partial X_{jr}}{\partial Y_r} = \sum_r X_{ir} \sum_r \frac{Y_r}{Y} \frac{\partial X_{jr}}{\partial Y_r}, \tag{12}$$

which implies identical Engel curves for all consumers It is immediately seen that (12) is sufficient but not necessary for (10) to hold An examination of higher-order terms reveals no further restrictions The calculations which are tedious and space consuming, but which involve no further points of principle, are left to the reader to work out

Even if individual consumer preference functions do conform to the condition imposed by (11), we must still face the fact that a move to a higher social behavior surface may involve some individuals gaining while others lose The caveats of the previous section still hold Nevertheless, there is an interesting, though still limited, interpretation which can be given to the aggregate equivalent variation

Suppose that we are considering a particular project costing the authorities X dollars per year Suppose also that this project generates an aggregate equivalent variation of Y dollars per year The key point is that the distribution of money income will be the same whether the project is undertaken or the equivalent variation is paid out Thus, the aggregate equivalent variation can be considered to be the opportunity cost of undertaking this project [2] If Y were less than X, it would be cheaper to pay individuals rather than to undertake the policy

It should be noted that this discussion has not involved the assumption of homotheticity The reason for this is that we are concerned with comparing alternative social states only with the current, existing situation Consider three social states the existing one A and two alternatives B and C The procedure described here enables us to calculate the equivalent variations necessary to achieve the same level of satisfaction as these two alternatives However, we do not require that the difference between these two numbers should also equal the equivalent variation associated with moving from B to C or vice versa Practical considerations require simply that we be able to rank alternatives with respect to the current situation

[2] This interpretation would seem to be in the spirit of Harberger's "third postulate" for applied welfare economics (1971, 1974, p 3)

5. A Final Point

As shown above, the proposed indicator is expressed in money terms This suggests a practical criterion for limiting the order of approximation being undertaken A cost of utility index Y/U can be constructed using current money income as the base for comparison Since index numbers are typically calculated to the nearest percentage, we could continue to add higher-order terms to the approximation until there are no further changes in the index While arbitrary, such a criterion is fairly realistic, since it is not likely that the underlying data being used can really be considered accurate beyond two or three digits

References

Harberger, A "Three Basic Postulates for Applied Welfare Economics " *J Econ Literature* 9, no 3 (September 1971) 785–97

——— *Taxation and Welfare* Boston Little, Brown, 1974

McKenzie, G , and Pearce, I F "Exact Measures of Welfare and the Cost of Living " *Rev Econ Studies* (1976), in press.

Pearce, I F *A Contribution to Demand Analysis* London Oxford Univ Press, 1964

The Effects of Property Taxes and Local Public Spending on Property Values: A Theoretical Comment

Bruce W Hamilton

Johns Hopkins University and Urban Institute

In a recent article in this *Journal*, Wallace Oates (1969) offered an empirical test of the Tiebout hypothesis Oates measured the importance of local fiscal variables, such as property tax rates and public school expenditure, in determining residential choice by observing the effect of these variables on residential property values in a cross-section of communities in northeastern New Jersey He of course corrected for other determinants of property values such as distance from New York City and average income of residents

Oates found property values to be positively correlated with public school expenditure and negatively correlated with the equalized local property tax rate Also, he showed that, when reasonable values of the variables are chosen, an increase in the property tax rate and a concomitant increase in public school expenditure give rise to only a negligible net change in property values These results, Oates claims, give strong support to the Tiebout hypothesis consumers do recognize fiscal differences between communities with unusually attractive features

The Oates paper leaves the impression that consumer responsiveness to local fiscal variables must necessarily, in a system of independent communities, lead to a correlation between fiscal variables and property values similar to that which he has observed in New Jersey But this is incorrect The Oates test rests on the assumption that a tax or benefit haven (collectively referred to here as a "fiscal haven")[1] will earn a rent, as reflected by higher housing prices In the long run, only goods produced

---

[1] Note that a fiscal haven cannot be created or maintained except by using some form of land-use restriction, such as zoning A low-tax-rate, high-public-expenditure community is a contradiction in terms unless the community has a higher than average tax *base* But the tax base is determined by residents' housing consumption In the absence of land-use restrictions, people with below-average demand for housing will move into fiscal havens, thus driving down the tax base and destroying the fiscal haven in the process I have examined the role of zoning in a Tiebout-type framework at length elsewhere (1972) George Peterson (1972) offers both theoretical and empirical evidence that land-use restrictions play a significant role in the system of communities examined by Oates

[*Journal of Political Economy*, 1976, vol 84, no 3]
© 1976 by The University of Chicago All rights reserved

under increasing cost can earn a rent. The extreme case, of course, is a good in perfectly inelastic supply, such as land 5 miles from Manhattan. Such land will earn a scarcity rent due to its access to New York. But I see no reason why a second community with a given set of fiscal characteristics cannot be created as cheaply as the first The scarcity of land is a problem common to all communities and should not influence the relative cost of creating various kinds of communities

The point now is quite simple Land at a given distance from the city center (more generally, the land containing a given "amount" of a nonreproducible, desirable characteristic) will, in equilibrium, be allocated among competing activities in such a manner that all such sites earn the same rent If one site earns more rent than a second identical site, the activity on the first site will expand at the expense of the activity occupying the second site until the rent discrepancy is eliminated

But *if all sites at a given distance from Manhattan earn the same rent* (i e , have the same value), then distance from Manhattan should explain all the variation in property values, and there will be nothing left to be explained by fiscal variables [2]

Oates's finding of a significant correlation between fiscal variables and property values must be due to one of the following two factors (1) a disequilibrium in which there is a temporary shortage of fiscal havens, or (2) persistent systematic differences in production functions either for raising revenue[3] or producing public services [4]

In the latter case, the consumer pays a site rent attributable to whatever natural or locational advantages the site may have, and he pays an additional rent in order to gain access to the nonreproducible efficiency of the local government's production This situation is completely analogous to the case of a single firm in an industry which is more efficient than its competitors As long as this situation prevails, the firm will earn a rent.

---

[2] This refers to the price of a "unit" of raw land, whereas the dependent variable in the Oates paper is the value of property consumed by the average homeowner in a community which includes the cost of the capital attached to a piece of land But since the price of capital is determined in a national market, it should not vary with local fiscal variables any more than the price of land (There are, of course, interregional differences in capital costs, but we are concerned here with interjurisdictional comparisons within a given urban area) Also, given the price of land and capital, the value of property consumed depends upon the amount purchased, which in turn presumably depends upon the income of the household So variables such as income, which are arguments in the demand function for housing, should be correlated with house values But fiscal variables, which in the long run, do not influence the prices of the inputs into housing should not in the long run be correlated with house value

[3] One argument in the production function for raising revenue is the community's industrial tax base (assuming the industry generates more revenue than cost), another is the community's access to state and federal revenue

[4] One important source of systematic interjurisdictional differences in production of public services may arise in public education It has been suggested that the efficiency of a public school system depends heavily upon who is being educated The children of middle-class families are more easily (and cheaply) taught than are the children of the poor. Thus a high-income community can produce education more efficiently than can a low-income community and is able to command a rent based on the nonreproducible efficiency of its production function

The disequilibrium explanation is also straightforward Due, say, to rising income levels or particular net migration flows in an urban area, a relative shortage of a given type of community occurs The immediate impact of this is a rise in property values in the scarce communities Thus, a rent difference will persist until property owners have shifted enough land into the scarce activity to eliminate the difference

It is not necessary that the communities in short supply be fiscal havens, although that happens to be the case in the Oates sample There could just as well be a shortage of low-expenditure communities, in which case property values in these communities would be bid up, and the public school expenditure variable in an Oates-type regression would have a negative, rather than a positive, coefficient This is completely analogous to the inevitable rise in the price of low-quality housing in the wake of slum-clearance programs

The paradoxical conclusion that "fiscal slums" might earn a rent is readily explained once the role of land-use restrictions is recalled Access to fiscal havens is restricted to those who are willing to purchase at least some minimum amount of property Fiscal slums, characterized by high tax rates and low public expenditure, are chosen by some (presumably the poor) instead of fiscal havens, characterized by low tax rates, high public expenditure, *and* stringent land-use restrictions Fiscal havens would undoubtedly be chosen by everybody if location in fiscal havens were not tied to minimum property consumption Since fiscal slums are demanded by some people, a reduced supply of sites in fiscal slums can cause the remaining sites to earn a rent in the short run Such a set of results is exactly opposite to what Oates found, but, as we have seen, completely consistent with consumer sensitivity to fiscal variables

We have seen that *any* set of coefficients on fiscal variables is consistent with fiscal sensitivity of consumers, it all depends on which activities are in short supply But there is only one set of coefficients which is consistent with consumer *insensitivity* to fiscal variables, namely, a complete lack of correlation between property values and fiscal variables [5] If consumers are insensitive to these variables, there is no way the variables can influence property values

The Oates test, then, does what he claims for it, it rejects the hypothesis that consumers ignore local fiscal variables when making their location decisions However, it should be pointed out that the Oates results do not follow *necessarily* from the Tiebout hypothesis and that there is no reason to expect his results to be reproducible using other samples. Indeed, Henry Pollakowski (1973) has attempted to reproduce the Oates results using data from the San Francisco Bay area, with mixed results. I expect that the reason for the difference between the Pollakowski results and the Oates results lies not in the superior intellect of northeastern New Jersey homeowners, but rather in differences (between New Jersey and the San Francisco Bay area) in important short-run supply variables (such as rates of net migration of various income classes), which were omitted from both studies

---

[5] Such a lack of correlation is also consistent with sensitivity to fiscal variables when the system is in long-run equilibrium So a lack of correlation between house value and local fiscal variables cannot be taken as a rejection of the Tiebout hypothesis

References

Hamilton, Bruce W. "The Impact of Zoning and Property Taxes on Urban Structure and Housing Markets." Ph.D. dissertation, Princeton Univ., 1972.

Oates, Wallace E "The Effects of Property Taxes and Local Public Spending on Property Values An Empirical Study of Tax Capitalization and the Tiebout Hypothesis" *J P E* 77, no 6 (November/December 1969) 957–70.

Peterson, George E. "The Use of Capitalization Effects to 'Test' the Tiebout Hypothesis." Working Paper no 1207-4, Urban Inst., 1972.

Pollakowski, Henry O. "The Effects of Property Taxes and Local Public Spending on Property Values A Comment and Further Results" *J.P.E* 81, no 4 (July/August 1973) 994–1003.

Book Reviews

Quantitative Models of Commodity Markets Edited by WALTER C LABYS. Cambridge, Mass Ballinger Publishing Co , 1975 Pp xxi+404 $15 00

This book, the first in a projected series on commodity modeling, is a collection of 11 studies that investigate important primary commodities, illustrate different approaches to quantitative market modeling, and employ recent developments in the field Eight of these studies were previously unpublished Introductory and concluding chapters on modeling methodology and possibilities for future research provide a frame of reference for the different studies There is also a bibliography of 241 market models for 66 commodities According to Labys, the book should inform model builders of new techniques, show economists how important markets work, and convince policymakers of the value of quantitative models I think the book provides a good survey of current market modeling that will interest econometric model builders But the book sheds little light on how the markets work or on current commodity problems

The market models for copper, coffee, cocoa, livestock, sugar, lauric oils, and fish take a supply and demand approach Linear programming is used to model the petroleum industry and grain-livestock trade in the European Economic Community Econometric techniques in these studies range from ordinary least squares to full information maximum likelihood—ordinary least squares and two-stage least-squares are most common Other quantitative techniques used are stochastic control, stochastic simulations, and systematic tests of models' predictions

The most interesting papers are a study of copper by Franklin Fisher, Paul Cootner, and Martin Baily and a study of sugar by Clifford Wymer The essay on copper estimates supply, demand, and stockholding for major producing and consuming nations and then investigates how government control of Chile's copper production could affect the industry's revenue Besides presenting a good market model that takes account of substitutes for copper, short-run, and long-run distinctions, and certain peculiarities of the world copper market, the paper employs advanced econometrics The essay on sugar presents an econometric technique for estimating continuous time models with discrete data Continuous time is often most appropriate or convenient for theoretical analysis of price adjustments and disequilibrium, so it is important to develop techniques for testing implications from such models The model of the world sugar market is used to illustrate the econometric technique and is not developed in much detail. Nevertheless, it includes the futures market for sugar and interaction between the spot and futures markets This approach could be useful for empirical work investigating the effects of futures trading

A striking feature of this volume is the nonfunctional role that many authors assign to market prices Yearly demands for gasoline are specified as zero elastic, and prices are determined by a markup over cost We are told that widely fluctuating cocoa prices cannot provide accurate signals for resource allocation.

dychotomy between the comparative static and the short-run dynamic approaches. Statistical purists might object to the lack of interequation constraints which arise from the production function. Finally, data purists might object to the authors' capacity utilization measure. All these critical points have some validity and undoubtedly they have already prompted several term papers or Ph.D dissertations But it seems most unlikely that such relatively minor changes in theoretical and statistical procedure would alter the authors' main message.

This study by Nadiri and Rosen is an important contribution to the development of factor demand models, not so much because of its positive results as because of its failure to generate satisfactory evidence of factor substitution in response to input price changes Their positive results, namely, that there are accelerator-type input demand functions with complex lags, are after all not unfamiliar from the literature, though their theoretical and statistical methods of obtaining them are novel and elegant. Factor substitutability, on the other hand, is an essential aspect of the conventional neoclassical theory of the firm Moreover, some other authors have recently claimed strong empirical support for this theory and even used factor substitutability as the basis for their policy recommendations Nadiri and Rosen's appropriate theoretical specification and careful empirical analysis have shown most convincingly that factor substitutability is *not* supported by their evidence and, therefore, it seems premature to use it as the basis for policy recommendations

The insufficient empirical support for factor substitution thus casts substantial doubt on the adequacy of the conventional neoclassical factor demand model as a suitable description of actual investment and employment behavior. How, then, should interested researchers proceed from here? In their final paragraph, Nadiri and Rosen suggest that research efforts should be devoted to first, adequate estimating techniques for dealing with serial correlation, second, improving the quality of the data and third, incorporating the supply side of the market into the model Upon reviewing the Nadiri-Rosen evidence it would appear that their third point deserves special emphasis. In a world with bottlenecks, delivery lags, and similar supply constraints it would surely be overly optimistic (indeed, naive) to expect that all observed levels of employment and capital differ from the values predicted by their demand functions by only randomly distributed errors Moreover, if standard capital demand functions are fitted to data which are generated artificially from plausible models with upward-sloping flow supply curves of capital, two biases emerge (1) the response of the demand for capital to changes in the price of capital is biased upward (from negative to positive), and (2) the lags are biased upward. These are, of course, precisely the biases that must be suspected in the Nadiri-Rosen results, and, hence, they constitute another reason for giving serious considerations to supply factors

Nadiri and Rosen deserve to be congratulated on this study, which meets the highest standards of careful theorizing and objective interpretation of the evidence. While it does not answer all our questions, it gives fairly clear directives to future researchers who, it is hoped, will adopt similarly high standards.

FRANK BRECHLING

Northwestern University

Economies of Scale in U.S. Electric Power Generation

Laurits R. Christensen and William H. Greene

University of Wisconsin

We estimate economies of scale for U.S. firms producing electric power. Cross-section data for 1955 and 1970 are analyzed using the translog cost function. We find that in 1955 there were significant scale economies available to nearly all firms. By 1970, however, the bulk of U.S. electricity generation was by firms operating in the essentially flat area of the average cost curve. We conclude that a small number of extremely large firms are not required for efficient production and that policies designed to promote competition in electric power generation cannot be faulted in terms of sacrificing economies of scale.

Rates charged for electric power in the United States have recently risen at historically unprecedented rates This has led to renewed scrutiny of proposals for reorganization of the U.S. electric power industry. One of the most interesting recent proposals has been made by Weiss (1975). He has suggested that firms in the industry be vertically disintegrated by divorcing the generation of electricity from its transmission and distribution. The motivation for this proposal is the belief that additional competition at the generation stage would result in downward pressure on rates. Competition, of course, requires that there be several potential suppliers for any particular load center. The crucial question pertaining to such a development is whether substantial scale economies would be sacrificed If so, the resulting inefficiency might offset the potential benefits of increased competition.

The regulated entity in the U.S electric power industry is the firm. Thus, assessment of the effects of reorganization requires information on

Associate professor of economics and graduate student, University of Wisconsin. We are grateful to Ernst Berndt, Dianne Cummings, Arthur Goldberger, Frank Gollop, and Leonard Weiss for helpful discussions and to the University of Wisconsin Graduate School for research support.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]

© 1976 by The University of Chicago. All rights reserved.

economies of scale for firms—as opposed to plants or individual generating units The purpose of this study is to provide such information. Our methodology is to model the structure of production in the industry using the neoclassical cost function approach. Pioneering work with this approach was done by Nerlove (1963) in his study of scale economies in electricity generation Recent developments in duality theory and functional form specification permit us to specify a much more general model than those available to Nerlove. In particular, our model imposes neither homotheticity nor unitary elasticities of substitution.

In estimation, it is essential to distinguish economies of scale from decreases in cost resulting from technical change The most convenient way to do so is to use cross-section data on firms which have had access to the same choices in plant design. It is also desirable, however, to gain some perspective on how scale economies have varied over time. To accomplish this, we have chosen to utilize two cross sections, the first for 1955 and the second for 1970. The 1955 sample facilitates comparison with the results of Nerlove, and the 1970 sample provides contemporary estimates.

Analysis of the two cross sections leads us to some important conclusions. First, we find that in 1955 there were significant scale economies available to nearly all firms By 1970, however, a large share of total electric power was generated by firms which had exhausted scale economies The explanation for this finding is that the firm size required to exhaust economies of scale increased by only 60 percent between 1955 and 1970, while actual output per firm approximately tripled

Second, most firms which generated electric power experienced substantial declines in average costs of production between 1955 and 1970. A standard explanation for this phenomenon is made in terms of the exploitation of economies of scale, principally the utilization of very large generator units Our findings do not support such an explanation. In fact, there was very little relationship between the degree of cost decline and the degree of expansion Many firms which had very low growth rates achieved cost reductions as impressive as those of the fastest-growing firms

Our primary finding, then, is that the U S electric power industry can be characterized by substantial scale economies at low levels of output. But the implied decreases in average cost diminish in importance for larger firms, resulting in an average cost curve which is very flat for a broad range of output. In 1955 most firms were producing on the low side of this range. By 1970, however, the bulk of U S electricity generation was by firms operating in the essentially flat area of the average cost curve. We conclude that a small number of extremely large firms are not required for efficient production and that policies to promote competition in electric power generation cannot be faulted in terms of sacrificing economies of scale.

I. The U.S. Electric Power Industry

The dominant mode of electricity generation in the United States is steam-driven turbines where the steam is produced by fossil-fuel-fired boilers. Water power has traditionally been an important mode as well, but a paucity of attractive dam sites and the tremendous growth of demand for electricity have combined to shrink its relative importance. Nuclear reactors may play a large future role, but their present contribution is still quite small. Internal-combustion engines are also employed but generally only in periods of peak demand. Table 1 indicates the mix of generator units in 1950 and 1970.

The technology and economics of the various forms of electricity generation are quite different. Given the dominant role of steam power, it is the logical starting point for any study which aspires to be relevant to policy decisions affecting the industry. Thus, we limit our attention to conventional steam-powered generators by analyzing only the costs of steam generation for each firm examined Steam generation accounts for more than 90 percent of total generation for nearly all of the firms in the sample We also limit our attention to investor-owned utilities with more than $1 million in annual revenues[1] These firms accounted for approximately 77 percent of the total power produced in the United States in 1970 The rest was generated by several federal projects and several hundred small firms, cooperatives, municipal districts, and state systems It would clearly be unwise to lump all of these diverse economic entities together in this study

Most observers agree that there exist economies of scale in electricity generation[2] There is disagreement, however, about the range over which economies persist. Some engineering estimates predict that economies of scale persist almost indefinitely For example, Hulbert (1969) claims that there are economies to be achieved for systems up to at least 25,000 megawatts.[3] In sharp contrast are the results of Johnston (1960) and Nerlove (1963), who found that scale economies were exhausted at relatively modest firm size Of course, many years have passed since their econometric estimates were made, so they are not necessarily inconsistent with contemporary engineering estimates.

In 1955 sales to ultimate consumers by privately owned electric power companies totaled 369 billion kilowatt-hours (kwh) By 1970 the total was 1,083 billion. Since the number of firms declined slightly, the output per firm increased approximately threefold. It is possible that changes in technology between 1955 and 1970 were great enough to have considerably expanded the firm size necessary to exploit scale economies. On the other

---

[1] Firms having revenues in excess of $2.5 and $1.0 million are designated by the Federal Power Commission as Class A and Class B, respectively.

[2] This paragraph follows closely the discussion of Weiss (1975, p. 148).

[3] This is almost twice the 1970 capacity of the nation's largest privately owned power system, the American Electric Power Company.

TABLE 1

PERCENTAGE OF TOTAL PRODUCTION BY TYPE OF PLANT*

Year	Steam	Hydroelectric	Internal Combustion	Nuclear
1950	69.8	29.1	1.1	0
1970	90.7	6.2	1.5	1.6

* The 1950 figures are from Caywood (1956) and refer to total U S production. The 1970 figures are from the Federal Power Commission (1971a) These figures are for investor-owned utilities only and neglect the large federal projects. Inclusion of these would increase the percentage of hydroelectric power somewhat. For example, TVA produces 6 percent of the nation's power, and 20 percent of its capacity is hydroelectric (*Forbes Magazine* 1975, p 25)

hand, if the previous econometric estimates are valid and if technological change has not dramatically changed the shape of the firm cost curve, the rapid growth of firms may have resulted in the exhaustion of economies of scale

In view of the disparate engineering and econometric estimates of the range of scale economies and the rapid growth of the industry since the econometric studies were performed, there is a clear need for further econometric analysis of the structure of production in the U S electric power industry.

II. Modeling the Structure of Production with a Neoclassical Cost Function

The recent application of duality theory to problems in economics has resulted in many useful results for the study of production and cost relationships.[4] A fundamental result is that, given certain regularity conditions, there exist cost and production functions which are dual to each other. The specification of a production function implies a particular cost function, and vice versa. Thus, the structure of production can be studied empirically using either a production function or a cost function; the choice should be made on statistical grounds. Direct estimation of the production function is attractive when the level of output is endogenous. Estimation of the cost function is more attractive, however, if the level of output is exogenous.

With minor exceptions, electric utilities are not allowed to choose the level of production to maximize profits They are required to supply all the electric power which is demanded at regulated prices. Entrepreneurial decisions are made with respect to the determination of the optimal levels of inputs.[5] Thus, the specification that input levels are endogenous variables and that the output level is an exogenous variable is plausible. Furthermore, since electric utilities compete with other industries for factors of production, the specification that input prices are exogenous variables is also reasonable. The cost function has as its arguments the level of output and the factor prices. Thus, estimation of the cost function is more attractive than direct estimation of the production function for studying the structure of production in the U.S. electric power industry. This was recognized by Nerlove (1963), and his study was one of the earliest applications of the principles of duality in modeling the structure of production.

---

[4] See Diewert (1974) for an excellent survey of the literature

[5] Whether or not regulated utilities actually do try to minimize costs has been the subject of considerable debate. Proponents of the Averch and Johnson (1962) hypothesis suggest, instead, that a utility faced with rate-of-return regulation will tend to over-

Recent developments in duality theory have enhanced the appeal of the cost function approach Every cost function implies a set of derived demand equations. Functional forms for cost functions have been developed which have two attractive features· they imply derived demand equations which are linear in the parameters, and at the same time they represent very general production structures, even though they cannot be derived from explicit production functions

We have chosen the translog cost function (Christensen, Jorgenson, and Lau 1971, 1973) for this study It places no a priori restrictions on substitution possibilities among the factors of production. Equally important, it allows scale economies to vary with the level of output. This feature is essential to enable the unit cost curve to attain the classical U shape The translog form is also attractive for this application because it contains all the forms estimated by Nerlove (1963) as special cases This allows us to perform statistical tests of the restrictions implied by Nerlove's models and to assess the impact that these restrictions have on estimates of scale economies

III. The Translog Cost Function

The translog cost function can be written

$$\left.\begin{aligned} \ln C = \alpha_0 &+ \alpha_Y \ln Y + \tfrac{1}{2} \gamma_{YY}(\ln Y)^2 \\ &+ \sum_i \alpha_i \ln P_i + \frac{1}{2} \sum_i \sum_j \gamma_{ij} \ln P_i \ln P_j \\ &+ \sum_i \gamma_{Yi} \ln Y \ln P_i, \end{aligned}\right\} \text{Model A}$$

---

capitalize Regarding electric utilities, some tentative support for the hypothesis is offered in the recent studies by Courville (1974), Spann (1974), and Petersen (1975). Breyer and MacAvoy (1974, p. 108) point out, however, that the activities of the electric utilities in the 1960s were actually contrary to predictions of the Averch-Johnson hypothesis Moreover, rapid increases in capital costs and the existence of considerable regulatory lag suggest that there was scant use of unnecessary capital in 1970—the year of chief interest for this study.

where $\gamma_{ij} = \gamma_{ji}$, C is total cost, Y is output, and the P_i's are the prices of the factor inputs. In order to correspond to a well-behaved production function, a cost function must be homogeneous of degree one in prices; that is, for a fixed level of output, total cost must increase proportionally when all prices increase proportionally. This implies the following relationships among the parameters

$$\sum_i \alpha_i = 1, \tag{1}$$

$$\sum_i \gamma_{Yi} = 0, \tag{2}$$

$$\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = \sum_i \sum_j \gamma_{ij} = 0 \tag{3}$$

A convenient feature of the cost function approach is that the derived demand functions for the factors of production can be easily computed by partially differentiating the cost function with respect to the factor prices, that is,

$$\partial C/\partial P_i = X_i \tag{4}$$

This result, known as Shephard's lemma (Shephard 1953), is conveniently expressed in logarithmic form for the translog cost function as

$$\partial \ln C/\partial \ln P_i = \frac{P_i X_i}{C} = S_i, \tag{5}$$

where S_i indicates the cost share of the ith-factor input The translog cost function yields the cost share equations

$$S_i = \alpha_i + \gamma_{Yi} \ln Y + \sum_j \gamma_{ij} \ln P_j \tag{6}$$

Uzawa (1962) has shown that Allen partial elasticities of substitution (Allen 1938) can be computed from the cost function by the formula

$$\sigma_{ij} = CC_{ij}/C_iC_j, \tag{7}$$

where subscripts on C indicate partial differentiation of the cost function with respect to factor prices For the translog cost function we have

$$\begin{aligned} \sigma_{ij} &= (\gamma_{ij} + S_iS_j)/S_iS_j, \\ \sigma_{ii} &= [\gamma_{ii} + S_i(S_i - 1)]/S_i^2, \qquad i \neq j \end{aligned} \tag{8}$$

The own-price elasticity of demand for the ith factor of production is

$$\eta_i = \sigma_{ii}S_i \tag{9}$$

Economies of scale are usually defined in terms of the relative increase in output resulting from a proportional increase in all inputs Hanoch

(1975) has pointed out that it is more appropriate to represent scale economies by the relationship between total cost and output along the expansion path—where input prices are constant and costs are minimized at every level of output. A natural way to express the extent of scale economies is as the proportional increase in cost resulting from a small proportional increase in the level of output, or the elasticity of total cost with respect to output. We will define scale economies (SCE) as unity minus this elasticity.

$$\mathrm{SCE} = 1 - \partial \ln C / \partial \ln Y \tag{10}$$

This results in positive numbers for positive scale economies and negative numbers for scale diseconomies Furthermore, SCE has a natural interpretation in percentage terms

The translog cost function does not constrain the structure of production to be homothetic, nor does it impose restrictions on the elasticities of substitution However, these restrictions can be tested statistically If any of the restrictions are valid, it is preferable to adopt the simplified model If not, it is of interest to investigate the impact of their imposition on the shape of the estimated cost curves

A cost function corresponds to a homothetic production structure if and only if the cost function can be written as a separable function in output and factor prices [6] A homothetic production structure is further restricted to be homogeneous if and only if the elasticity of cost with respect to output is constant For the translog cost function, the homotheticity and homogeneity restrictions are, respectively,

$$\gamma_{Yi} = 0 \tag{11}$$

and

$$\gamma_{Yi} = 0, \qquad \gamma_{YY} = 0 \tag{12}$$

The elasticities of substitution can all be restricted to unity by eliminating the second-order terms in the prices from the translog cost function. Thus, the unitary elasticity restrictions are

$$\gamma_{ij} = 0 \tag{13}$$

We consider five models in addition to the translog Model A Model B imposes homotheticity Model C imposes homogeneity. Models D, E, and F correspond to Models A, B, and C, respectively, with unitary elasticities of substitution imposed in each case The formulas for SCE for Models A–F are presented in table 2

---

[6] See Diewert (1974) for formal statements and derivations of the restrictions for homotheticity and homogeneity

TABLE 2

SCALE ECONOMIES FOR MODELS A–F

$\text{SCE (A)} = 1 - (\alpha_Y + \gamma_{YY} \ln Y + \Sigma_i \gamma_{Yi} \ln P_i)$

$\text{SCE (B)} = 1 - (\alpha_Y + \gamma_{YY} \ln Y)$

$\text{SCE (C)} = 1 - \alpha_Y$

$\text{SCE (D)} = 1 - (\alpha_Y + \gamma_{YY} \ln Y + \Sigma_i \gamma_{Yi} \ln P_i)$

$\text{SCE (E)} = 1 - (\alpha_Y + \gamma_{YY} \ln Y)$

$\text{SCE (F)} = 1 - \alpha_Y$

IV. Estimation Procedures

It is feasible to estimate the parameters of each cost function using ordinary least squares. This technique, used by Nerlove (1963), is certainly attractive from the point of view of simplicity. However, it neglects the additional information contained in the cost share equations (6), which are also easily estimable. Furthermore, even for a modest number of factors of production, the translog cost function has a large number of regressors which, aside from the terms involving output, do not vary greatly across firms. Hence, multicollinearity may be a problem, resulting in imprecise parameter estimates.

An alternative estimation procedure, followed by Berndt and Wood (1975), is to estimate the cost shares (6) as a multivariate regression system. This procedure was satisfactory for their purposes, since they treated constant returns to scale as a maintained hypothesis ($\gamma_{Yi} = 0$, $\gamma_{YY} = 0$, $\alpha_Y = 1$). Thus, their cost share equations included all the parameters of the cost function except for the neutral shift parameter α_0, and no economic information was lost by not including the cost function in the estimation procedure. This approach is not satisfactory, however, when constant returns are not imposed, since the crucial parameters in the SCE formulas, α_Y and γ_{YY}, appear only in the cost function.

We conclude that the optimal procedure is to jointly estimate the cost function and the cost share equations as a multivariate regression system. Including the cost share equations in the estimation procedure has the effect of adding many additional degrees of freedom without adding any unrestricted regression coefficients. This will result in more efficient parameter estimates than would be obtained by applying ordinary least squares to the cost function alone.

We specify additive disturbances for each of the share equations and for the cost function. Since the cost share equations are derived by differentiation, they do not contain the disturbance term from the cost function. We assume that the disturbances have a joint normal distribution. Following Zellner (1962), we allow nonzero correlations for a particular firm but impose zero correlations across firms. However, his proposed estimation procedure is not operational for our model. The estimated disturbance covariance matrix required to implement Zellner's procedure is singular because the disturbances on the share equations must sum to zero for each firm The Zellner procedure can be made operational by deleting one of the share equations from the system However, the estimates so obtained will not be invariant to which equation is deleted

Barten (1969) has shown that maximum-likelihood estimates of a system of share equations with one equation deleted are invariant to which equation is dropped A straightforward extension of his result covers our multivariate system, which includes the cost function with the share equations. Kmenta and Gilbert (1968) and Dhrymes (1970) have shown that iteration of the Zellner estimation procedure until convergence results in maximum-likelihood estimates Iterating the Zellner procedure is a computationally efficient method for obtaining maximum-likelihood estimates, and is the procedure which we employ here

Since we obtain maximum-likelihood estimates, we are able to test hypotheses such as homotheticity by using the likelihood ratio test Denoting the determinants of the unrestricted and restricted estimates of the disturbance covariance matrix by $|\hat{\Omega}_U|$ and $|\hat{\Omega}_R|$, respectively, we can write the likelihood ratio

$$\lambda = (|\hat{\Omega}_R|/|\hat{\Omega}_U|)^{-T/2}, \tag{14}$$

where T is the number of firms We test hypotheses using the fact that $-2 \ln \lambda$ is distributed asymptotically as chi-squared with degrees of freedom equal to the number of independent restrictions being imposed

V. Data

Since fuel accounts for a very large share (approximately 65 percent) of the cost of electricity generated by steam power, it is essential to include fuel in the model as a factor of production Intermediate inputs other than fuel are relatively unimportant in the generation stage of electricity production Thus, it is realistic to model the structure of production with capital, labor, and fuel as inputs Hence, we require prices and cost shares for these three inputs in addition to the quantity and cost of total output

To provide a link with Nerlove's (1963) results, we initially estimated our models using the same data as Nerlove—a 1955 cross section of Class A and B firms Nerlove did not include the cost share equations in his estimation procedure and therefore did not construct the cost share data We reconstructed the cost components using Nerlove's primary source (Federal Power Commission 1956)

Weiss (1975, p 148) has pointed out that Nerlove may have considerably underestimated the extent of scale economies by treating members of holding companies as if they were individual firms. Since the holding company, rather than its subsidiaries, is the proper unit of observation, we revised Nerlove's data by combining these firms where appropriate. This reduces the number of observations in the sample from 145 to 124.

With two exceptions, we used Nerlove's procedures to construct a 1970 cross section of 114 firms and holding companies First, for the price of fuel Nerlove used the statewide average from the state in which the firm was principally located The Federal Power Commission (1971*b*) provides a plant-by-plant summary of fuel purchases by type of fuel (coal, oil, and gas), by BTU equivalents, and by prices paid per million BTU for each type of fuel By simply converting all fuel purchases at all sites to BTU equivalents, we were able to average these prices, using as weights the contribution of each fuel purchase to total BTU consumption This procedure yields a firm-specific price for fuel Second, for the price of labor in 29 states Nerlove used statewide averages of wages for utility workers In the 19 remaining states not even the statewide average wage for utility workers was available For these states Nerlove estimated utility workers' wages from the regression relationship between utility workers' wages and all production workers' wages in the other 29 states Unlike Nerlove's source (FPC 1956), FPC (1971*a*) provides enough information to construct a yearly labor cost for each firm. We computed the total labor cost for each firm as the sum of total salaries and wages paid and employee pensions and benefits To arrive at a yearly price of labor, we divided this by the number of full-time employees plus one-half the number of part-time employees While this is still not specific to steam generation, it represents a clear improvement over Nerlove's procedure

VI. Empirical Results

a) Cost Functions for the U S Electric Power Industry

We present results for cost function Models A–F for three data sets The data set used by Nerlove is denoted 1955I. Nerlove's data set with firms combined into holding companies is denoted 1955II Our new data set is denoted 1970. Capital, labor, and fuel inputs are, respectively, denoted by K, L, and F subscripts

All six models were estimated for all three data sets The linear homogeneity in factor prices constraints (1, 2, 3) were imposed throughout. Additional regularity conditions which the cost functions must satisfy in order to correspond to well-behaved production structures are monotonicity and convexity in factor prices Sufficient conditions for these to hold are positive fitted cost shares and negative definiteness of the bordered Hessian of the cost function. These conditions are met at every observation for all of the models which we estimated; hence, we may conclude that the estimated cost functions represent well-behaved production structures.

TABLE 3

COST FUNCTION PARAMETER ESTIMATES, 1955I DATA
(t-RATIOS IN PARENTHESES)

PARAMETERS	MODEL A	B	C	D	E	F
α_0	8.841 (37.21)	8.037 (31.77)	6.761 (31.70)	8.366 (47.29)	7.866 (47.90)	6.542 (62.13)
α_Y	0.252 (4.34)	0.335 (5.90)	0.810 (55.88)	0.223 (3.75)	0.300 (5.14)	0.797 (53.43)
α_K	0.076 (0.91)	0.122 (1.31)	0.122 (1.34)	0.449 (12.98)	0.426 (44.18)	0.426 (44.09)
α_L	0.365 (4.95)	0.117 (1.29)	0.120 (1.36)	0.210 (20.44)	0.106 (27.88)	0.106 (27.88)
α_F	0.559 (9.44)	0.761 (13.93)	0.758 (14.02)	0.341 (9.55)	0.468 (44.78)	0.468 (44.72)
γ_{YY}	0.079 (7.96)	0.079 (7.97)		0.083 (8.18)	0.083 (8.24)	
γ_{YK}	−0.008 (−1.69)			−0.004 (−0.693)		
γ_{YL}	−0.017 (−11.08)			−0.016 (−10.49)		
γ_{YF}	0.025 (5.58)			0.020 (3.69)		
γ_{KK}	0.175 (6.31)	0.155 (5.34)	0.154 (6.20)			
γ_{LL}	0.040 (2.34)	0.033 (0.11)	0.004 (0.15)			
γ_{FF}	0.167 (6.95)	0.154 (6.07)	0.154 (6.58)			
γ_{KL}	−0.024 (−1.51)	−0.022 (−0.054)	−0.002 (−0.073)			
γ_{KF}	−0.151 (−6.43)	−0.153 (−6.52)	−0.152 (−6.77)			
γ_{LF}	−0.016 (−1.94)	−0.001 (−0.113)	−0.002 (−0.169)			
Restrictions	None	(2)	(3)	(3)	(5)	(6)
Log det $\hat{\Omega}$	2.261	2.903	3.231	2.550	3.163	3.497

The complete set of parameter estimates for 1955I is presented in table 3. To conserve space, we present in table 4 only the Model A parameter estimates for the 1955II and 1970 data sets. The t-ratios for the nonhomotheticity parameters (γ_{Y_i})and the substitution parameters (γ_{ij}) for Model A suggest that neither the homotheticity hypothesis nor the unitary elasticities of substitution hypothesis is consistent with any of the data sets. This suggestion is confirmed by the likelihood ratio statistics given in table 5 for all three data sets.[7] All of the restricted models are

---

[7] The estimates for Models E and F differ from those given by Nerlove (1963), since he used the OLS estimation procedure. We also computed the OLS estimates of the cost function parameters. They turn out to have generally larger standard errors than the efficient estimates. With these estimates we are not able to reject the homotheticity hypothesis or unitary elasticities of substitution hypothesis for any of the three data sets.

TABLE 4

COST FUNCTION PARAMETER ESTIMATES FOR MODEL A, 1955II AND 1970 DATA (t-RATIOS IN PARENTHESES)

Parameter	1955II	1970	Parameter	1955II	1970
α_0	8.412 (31.52)	7.14 (32.45)	γ_{YF}	0.024 (5.14)	0.021 (6.64)
α_Y	0.386 (6.22)	0.587 (20.87)	γ_{KK}	0.175 (5.51)	0.118 (6.17)
α_K	0.094 (0.94)	0.208 (2.95)	γ_{LL}	0.038 (2.03)	0.081 (5.00)
α_L	0.348 (4.21)	−0.151 (−1.85)	γ_{FF}	0.176 (6.83)	0.178 (10.79)
α_F	0.558 (8.57)	0.943 (14.64)	γ_{KL}	−0.018 (−1.01)	−0.011 (−0.749)
γ_{YY}	0.059 (5.76)	0.049 (12.94)	γ_{KF}	−0.156 (−6.05)	−0.107 (−7.48)
γ_{YK}	−0.008 (−1.79)	−0.003 (−1.23)	γ_{LF}	−0.020 (−2.08)	−0.070 (−6.30)
γ_{YL}	−0.016 (−10.10)	−0.018 (−8.25)			

TABLE 5

TEST STATISTICS FOR HOMOTHETICITY, HOMOGENEITY, AND UNITARY ELASTICITIES OF SUBSTITUTION

	HYPOTHESIS (RESTRICTION ON MODEL A)				
	Homotheticity	Homogeneity	Unitary Elasticities of Substitution	Homotheticity and Unitary Elasticities of Substitution	Homogeneity and Unitary Elasticities of Substitution (Cobb-Douglas)
Number of restrictions	2	3	3	5	6
Critical χ^2 (1%)	9.21	11.35	11.35	15.09	16.81
χ^2 1955I	93.09	140.51	41.91	130.79	179.22
χ^2 1955II	78.22	102.27	43.92	112.84	158.89
χ^2 1970	57.91	157.46	94.95	137.77	230.70

rejected at reasonable levels of significance in all three data sets. Thus, it is clear that a model which allows nonhomotheticity and nonunitary elasticities of substitution is required to adequately represent the structure of production for U.S. firms generating electric power.

It has often been suggested that substitution possibilities in electric power generation are scant. For example, Komiya (1962) argued in favor of modeling the structure of production with a fixed-coefficients technology, at least at the plant level. On the other hand, Nerlove (1963, pp. 173–74) argued that there may be significant substitution possibilities at the firm level. To provide some evidence on this question, we averaged

TABLE 6

ESTIMATED ELASTICITIES OF SUBSTITUTION AND OWN-PRICE ELASTICITIES

Elasticities of Substitution	Capital-Labor	Capital-Fuel	Labor-Fuel
1955I	.411	.223	650
1955II	.547	.195	.574
1970	.639	218	.165
Own-Price Elasticities	**Capital**	**Labor**	**Fuel**
1955I	– 159	–.499	– 170
1955II	– 163	–.520	– 146
1970	– 238	–.229	– 086

over firms the elasticities of substitution and own-price elasticities implicit in our cost function estimates The results for all three data sets are presented in table 6

b) *Economies of Scale*

An estimate of scale economies can be derived for each firm by evaluating the formulas in table 2 at the observed level of output and factor prices Following Nerlove, we partitioned the sample into five groups of firms according to output, in table 7 we present the 1955I estimates of scale economies for the firm with the median output in each group. The *t*-ratios given with the estimates use firm-specific standard errors, conditional on observed output and prices The estimates of scale economies for the homogeneous models, C and F, are constant for all levels of output. The other four models, which allow scale economies to vary with output, confirm our expectation that scale economies diminish as firm size increases. Although the imposition of unitary elasticities of substitution leads to a statistically significant loss of fit, the impact on estimated scale economies is not large The estimates for Models D, E, and F are roughly the same as those for A, B, and C, respectively However, the effect of the homotheticity constraint is much more pronounced. The estimates of scale economies for Models B and E, the homothetic models, are considerably less than those for Models A and D.

The homogeneous Models C and F give the erroneous impression that there are significant economies of scale for firms of any size. The nonhomogeneous models, however, indicate that scale economies are exhausted well within the sample output range The median firm in group 5, which had output less than one-third that of the largest firm, shows no significant scale economies.

TABLE 7

ESTIMATED SCALE ECONOMIES UNDER VARIOUS SPECIFICATIONS OF TECHNOLOGY
(*t*-RATIOS IN PARENTHESES)

	Size Group				
	1	2	3	4	5
Output (million kwh)	43	338	1,109	2,226	5,819
			1955I		
Model					
Homogeneous.					
F	203 (13 61)	203 (13.61)	.203 (13 61)	.203 (13 61)	.203 (13.61)
C	190 (13 11)	190 (13 11)	190 (13.11)	.190 (13.11)	.190 (13.11)
Homothetic					
E	.388 (17 00)	216 (16 90)	.117 (6 28)	059 (2 43)	−.020 (−0 62)
B	369 (16.62)	208 (16 66)	113 (6 20)	059 (2 44)	− 017 (−0 53)
Nonhomothetic					
D	418 (18 00)	258 (18.53)	153 (7 94)	.096 (3.83)	.026 (0 77)
A	408 (17.88)	.258 (18 44)	.157 (8 25)	104 (4.16)	.040 (1.20)
			1955II		
Model A	351 (13 66)	243 (15 67)	167 (8 68)	27 (5.15)	076 (2 26)

It is of interest to compare our estimates of scale economies with those obtained by Nerlove using ordinary least squares.[8] For Model F our estimate is 203, versus Nerlove's estimate of .281. Nerlove's estimates for Model E are inconsistent with his data; apparently he used the formula $(\alpha_Y + \frac{1}{2}\gamma_{YY} \log Y)^{-1}$ instead of $(\alpha_Y + \gamma_{YY} \log Y)^{-1}$ as implied by his model, to compute returns to scale. Table 8 presents our estimates along with Nerlove's corrected and erroneous estimate for Model E. The pattern we observe is that OLS leads to an increase in the estimates of returns to scale at low levels of output and a decrease at high levels. The same pattern emerges from the OLS estimates of Models A, B, and D.

A convenient way to summarize scale economies is to present the average cost curve facing a typical firm. The cost curve can be derived by evaluating the average cost function for a range of outputs while holding

---

[8] Nerlove computed returns to scale as the sum of the output elasticities. This is the inverse of one minus our estimate of scale economies. We used this transformation of Nerlove's estimates for comparability with our own. Also, Nerlove used common logarithms in his calculations while we used natural logarithms in ours.

TABLE 8

ESTIMATED SCALE ECONOMIES FOR MODEL E (1955I DATA)

	OUTPUT (MILLION kwh)				
	43	338	1,109	2,226	5,819
Christensen-Greene ..	388	.216	.117	.059	−.020
Nerlove (corrected) ..	.465	.257	.136	.065	−.032
Nerlove (erroneous) .	.658	.554	.492	457	408

the factor prices fixed at the sample means.[9] These curves are plotted for Models A–F in figure 1 (for 1955I). The size distribution of firms is presented beneath the cost curves Consistent with the figures in table 7, Models C and F have continuously declining average costs, while the others all have the classical U shape Dotted arrows indicate the minimum average cost corresponding to Models A, B, D, and E. The imposition of unitary elasticities of substitution has some effect on the shape of the curves, and imposition of homotheticity changes the shape of the cost curve substantially.

Figure 1 reveals that the cost curve for Model A is not well approximated by those from the restricted models Scale economies persist at outputs considerably above those indicated by Models D, B, and E. Even so, a substantial proportion of electric power appears to have been generated by firms operating in the essentially flat area of the cost curve, and scale economies appear to be exhausted by firms well within the size range of the sample

We found that Models B, C, D, E, and F can be rejected both on statistical grounds and because they do not provide good approximations to the unit cost curve from Model A We henceforth limit our attention to Model A in studying the 1955II and 1970 data sets

Weiss (1975) has suggested that Nerlove (1963) may have understated scale economies by treating members of holding companies as individual firms In our 1955II data set, observations for firms were combined when they were members of holding companies It turns out that the five firms in table 7 remained as observations, since they were not subsidiary firms in holding companies It provides a useful comparison to compute the estimates of scale economies for these five firms The estimates, presented for Model A in the last line of table 7, show that scale economies were less at lower levels of output—but persisted over a wider range of output. This finding is also apparent from the 1955II unit cost curve in figure 2. The difference between the 1955I and 1955II unit cost curves in figure 2 is

---

[9] The slope of the unit cost curve is sufficient to infer the presence of scale economies, since SCE = 1 − (marginal cost/average cost) Thus, declining unit costs indicate economies of scale and rising unit costs indicate diseconomies.

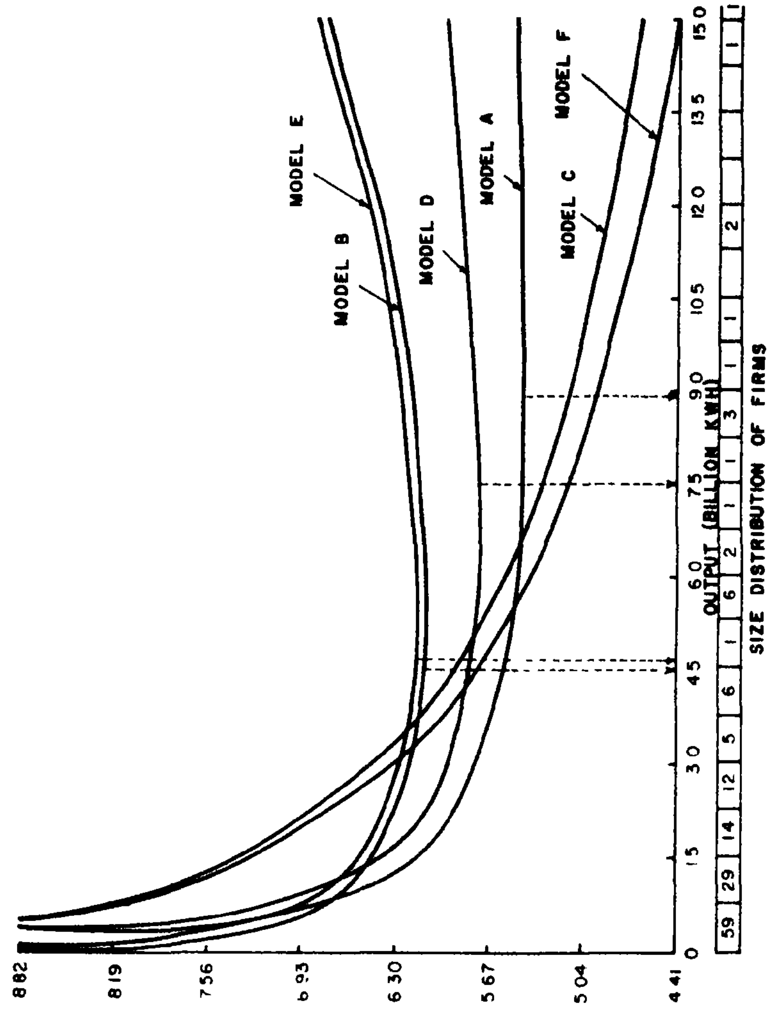
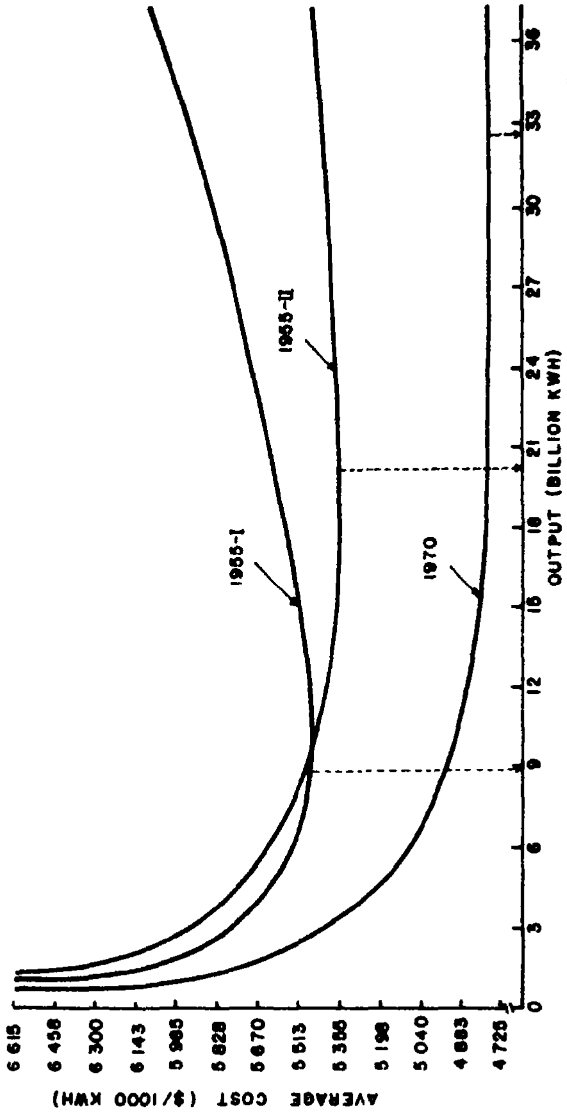


FIG 1 —Average cost curves, 1955I data (dotted arrow indicates minimum average cost)

striking. The 1955II curve is relatively flat over a much broader range of output than the 1955I curve In addition, the range of decreasing costs is more than double that of the 1955I curve. The size distribution of firms in figure 2 indicates that no firm had exhausted scale economies in 1955 Thus, Weiss's conjecture that the 1955I data set yields an understatement of scale economies is supported by our results.

Between 1955 and 1970, sales of electricity to ultimate consumers by privately owned producers grew from 369 to 1,083 billion kwh. Table 9



1955-I	88	26	11	7	3	4	2	2		1		1													
1955-II	76	15	8	8	3	7	2	2		2		1													
1970	26	22	7	10	9	1	6	4	3	2	1	4	1	1	1	2	2	1	2	1	1	1	2	1	3

SIZE DISTRIBUTION OF FIRMS

FIG. 2 —Average cost curves, Model A (dotted arrow indicates minimum average cost)

illustrates the pattern of growth for a sample of firms Results from fitting Model A with our 1970 data set are indicated in table 9 and figure 2. The last two columns of table 9 show that firms typically had lower scale economies available in 1970 than in 1955 The reasons for this are clear from figure 2. The average firm tripled its output level between 1955 and 1970, but the shape of the cost curve changed very little. In particular,

TABLE 9

ESTIMATES OF SCALE ECONOMIES FOR SELECTED FIRMS
(FIRMS ORDERED BY 1970 OUTPUT)

COMPANY	OUTPUT (MILLION kwh)		AVERAGE COST ($/1,000 kwh)		SCE		
	1955	1970	1955	1970	1955I	1955II	1970
Newport Electric	68	50	11 69	10 75	379	332	.324
Community Public Service	63	183	8 73	7 03	.413	363	247
United Gas Improvement	235	467	10 39	8 44	442	.267	.216
St Joseph Light & Power	253	938	8 03	5.45	284	263	.181
Iowa Southern Utilities	299	1,328	11 36	6 07	265	248	160
Missouri Public Service	209	1,886	9 26	5 47	299	274	143
Rochester Gas & Electric	1,156	2,020	7 66	8 89	158	167	136
Iowa Electric Light & Power	1,166	2,445	7 37	5 37	158	167	133
Central La Gas & Electric	353	2,689	7 53	5 54	253	263	127
Wisconsin Public Service	1,122	3,571	6 90	6 02	162	170	103
Atlantic City Electric	1,291	4,187	7 96	7 00	150	161	094
Central Illinois Public Service	2,304	5,316	5 48	4 43	110	133	097
Kansas Gas & Electric	1,668	5,785	5 08	3 36	136	153	094
Northern Indiana Public Service	1,137	6,837	7 01	4 96	167	174	079
Indianapolis Power and Light	2,341	7,484	5 51	3 94	112	135	080
Oklahoma Gas & Electric	2,353	10,149	4 94	3 01	119	142	066
Niagara Mohawk Power	8,787	11,667	5 47	6 40	– 003	047	049
Potomac Electric Power	3,538	13,846	6 79	6 95	070	102	037
Gulf States Utilities	2,507	17,875	3.85	3 27	121	146	036
Virginia Electric Power	5,277	23,217	6 12	4 85	037	077	015
Consolidated Edison	14,359	29,613	9 71	8 43	– 041	019	– 003
Detroit Edison	11,796	30,958	6 19	6 05	– 026	030	– 004
Duke Power	9,956	34,212	4 51	4 84	– 013	038	– 012
Commonwealth Edison	19,170	46,871	7 10	5 43	– 035	024	– 014
Southern	13,702	53,918	4 74	4 30	– 056	008	– 028

the bottom of the cost curve shifted out by a relatively modest 60 percent (from 20 to 32 billion kwh) Thus, most firms were operating in a portion of the 1970 cost curve which was flatter than their point of operation on the 1955II cost curve In effect, unexploited economies of scale were much smaller in 1970 than in 1955

Up to this point, we have spoken rather loosely about "flat" or "relatively flat" portions of the average cost curves We now give some precision to this notion by defining the "flat" portion of the curve as that part which exhibits no statistically significant economies or diseconomies of scale. We choose the 5 percent significance level and allocate 2½ percent to each tail for SCE Thus, any point on the average cost curve with a corresponding SCE which is less than 1 96 times its standard error is considered to be in the "flat" region of the curve. When a firm reaches a level of output in the "flat" region, we say that it has no statistically significant unexploited scale economies The "flat" regions of the cost curves in figure 2 are indicated in table 10 along with the number of

TABLE 10

RANGES OF SIGNIFICANT SCALE ECONOMIES

	1955I	1955II	1970
Bounds of the region with no significant economies or diseconomies (billion kwh):			
Lower	4,400	9,700	19,800
Upper	34,000	57,000	67,100
Significant scale economies			
Number of firms .	119	118	97
Proportion of total output	377	741	487
No significant economies or diseconomies·			
Number of firms	26	6	16
Proportion of total output	623	259	446
Significant scale diseconomies			
Number of firms	0	0	1
Proportion of total output	0	0	067

firms and proportion of output produced in each of three regions (1) statistically significant scale economies, (2) no significant scale economies or diseconomies, and (3) statistically significant scale diseconomies The changes in industry production are dramatic In 1955, 74 1 percent of steam-generated electric power was produced by firms with significant unexploited scale economies By 1970, the figure had fallen to 48 7 percent Not only was a large portion of all power produced in the flat region of the cost curve, but a nontrivial amount (6 7 percent) was produced by one firm (the American Electric Power Company) which showed significant diseconomies of scale

VII. Concluding Remarks

In figure 3 we display the 1970 average cost curve over the full range of output observations In addition, we include on the figure solid vertical arrows to indicate the bounds of the "flat" portion of the curve This demonstrates the large range of output which can be interpreted as exhibiting constant returns to scale Identification of this range provides an important tool for analyzing proposals to restructure the electric power industry. The aggregate cost of producing electric power is clearly related to the proportion produced by firms operating in the "flat" area of the cost curve.

Table 9 and figure 2 indicate that the cost of producing electric power fell absolutely during the 1955–70 period. This phenomenon is often attributed to the exploitation of scale economies or to scale-related technical change. Such explanations are contradicted by figure 2. To a

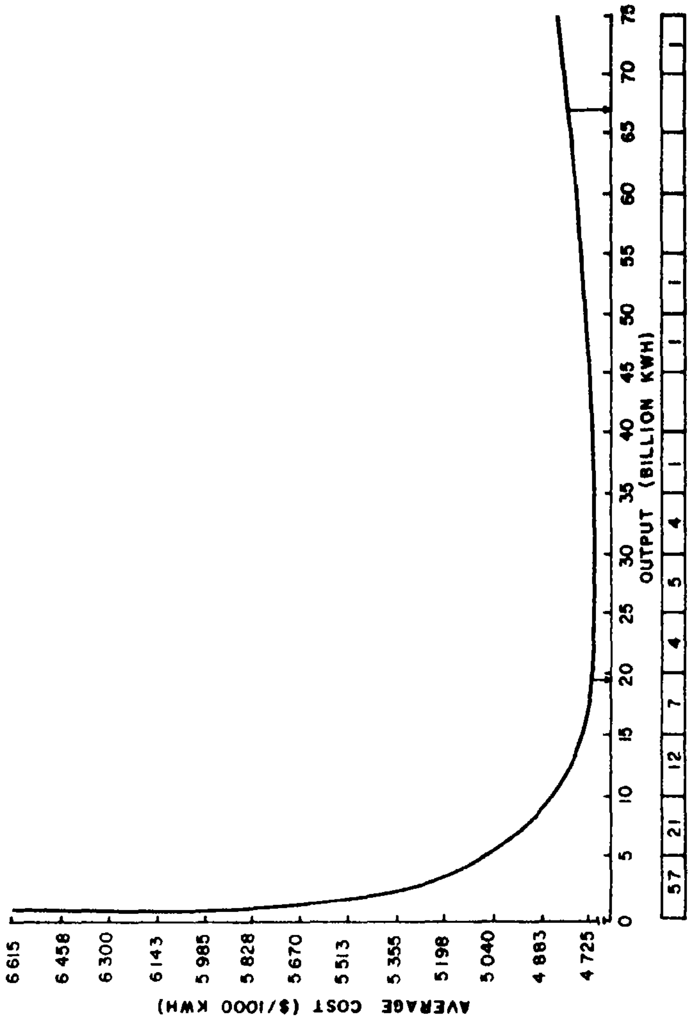


FIG 3 —Average cost curves, Model A, 1970 (arrows indicate region of no significant economies or diseconomies of scale)

close approximation, the 1970 cost curve is simply a downward vertical displacement of the 1955II curve This implies that technical change unrelated to increases in scale deserves the primary attribution for declines in the cost of production This is corroborated by the fact that there is little correlation between the degree of cost reduction and the rate of growth of individual firms It is true that many firms did exploit scale economies between 1955 and 1970 However, for the typical firm, movements along the cost curve resulted in cost declines which were small relative to the downward displacement of the entire curve

Hughes (1971) and others have discussed the desirability of rationalizing the U.S. electric power industry—reducing the number of firms or planning units. Our cost functions yield estimates of the potential cost savings from rationalization. In 1970 if the combined output of all the firms in our sample had been produced at the minimum point on the average cost curve (0.473¢ per kilowatt hour), the total cost of production would have been \$175.1 million less[10] The estimated cost decline is 3 2 percent of observed costs Of the estimated savings, \$83 8 million (47 9 percent) is for labor ,\$49 6 million (28 3 percent) for capital, and \$41 7 million (23.8 percent) for fuel Production of the observed output by the 114 firms in our sample would have required 33 firms of optimal size. The composition of the cost decline differs considerably from the average cost shares (14 percent for labor, 23 percent for capital, and 63 percent for fuel) The difference results from the nonhomotheticity of the production process

References

Allen, R G D *Mathematical Analysis for Economists* London Macmillan, 1938.

Averch, H , and Johnson, L L "Behavior of the Firm under Regulatory Constraint " *A E R* 52 (December 1962) 1052–69

Barten, A P. "Maximum Likelihood Estimation of a Complete System of Demand Equations " *European Econ Rev* 1 (Fall 1969), 7–73

Berndt, E R , and Wood, D O "Technology, Prices, and the Derived Demand for Energy." *Rev Econ and Statis* 57, no 3 (August 1975) 259–68

Breyer, S G , and MacAvoy, P *Energy Regulation by the Federal Power Commission* Washington Brookings Inst , 1974

Caywood, R E *Electric Utility Rate Economics* New York McGraw-Hill, 1956

Christensen, L R , Jorgenson, D W , and Lau, L J "Conjugate Duality and the Transcendental Logarithmic Production Function" (abstract). *Econometrica* 39 (July 1971) 255–56

——— "Transcendental Logarithmic Production Frontiers " *Rev Econ. and Statis* 55 (February 1973) 28–45

Courville, L "Regulation and Efficiency in the Electric Utility Industry " *Bell J Econ and Management Sci* 5 (Spring 1974) 53–74

Dhrymes, P J "Equivalence of Iterative Aitken and Maximum Likelihood Estimates for a System of Regression Equations " Unpublished Univ Pennsylvania, 1970

Diewert, W E "Applications of Duality Theory " In *Frontiers of Quantitative Economics*, edited by M D Intriligator and D A Kendrick Amsterdam North-Holland, 1974

Federal Power Commission *Statistics of Privately Owned Electric Utilities in the United States, Classes A and B, 1955 Supplement* Washington Government Printing Office, 1956.

——— *Statistics of Privately Owned Electric Utilities in the United States, Classes A and B, 1970 Supplement* Washington Government Printing Office, 1971. (*a*)

---

[10] This figure is arrived at by moving all production along the fitted average cost curve to the minimum point

———. *Steam Electric Plant Construction Costs and Annual Production Expenses, 1970 Supplement.* Washington: Government Printing Office, 1971. (*b*)

Forbes Magazine. "Tennessee Valley Authority: The Yardstick with Less than 36 Inches" (April 1, 1975), pp. 24–28.

Hanoch, G. "The Elasticity of Scale and the Shape of Average Costs." *A.E.R.* 65, 3 (June 1975) 492–97.

Hughes, W. R. "Scale Frontiers in Electric Power." In *Technological Change in Regulated Industries,* edited by William M Capron Washington: Brookings Inst., 1971.

Hulbert, G C "Power Generation in the 1970's." In *1969 Future Power Forum. Perspective for the 70's* New York. Westinghouse, 1969

Johnston, J J *Statistical Cost Analysis.* New York. McGraw-Hill, 1960.

Kmenta, J , and Gilbert, R. F. "Small Sample Properties of Alternative Estimators of Seemingly Unrelated Regressions." *J American Statis. Assoc.* 63 (December 1968) 1180–1200

Komiya, R "Technical Progress and the Production Function in the United States Steam Power Industry " *Rev. Econ and Statis* 44, no. 2 (May 1962)· 156–67

Nerlove, M "Returns to Scale in Electricity Supply." In *Measurement in Economics —Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld,* edited by Carl F Christ. Stanford, Calif Stanford Univ Press, 1963.

Petersen, H. C "An Empirical Test of Regulatory Effects." *Bell J Econ.* 6, no 1 (Spring 1975) 111–26

Shephard, R W *Cost and Production Functions* Princeton, N J Princeton Univ Press, 1953

Spann, R M "Rate of Return Regulation and Efficiency in Production An Empirical Test of the Averch-Johnson Thesis." *Bell J Econ and Management Sci* 5, no. 1 (Spring 1974) 38–52

Uzawa, H "Production Functions with Constant Elasticities of Substitution " *Rev Econ and Statis* 44, no 4 (October 1962) 291–99.

Weiss, L W "Antitrust in the Electric Power Industry " In *Promoting Competition in Regulated Markets,* edited by Almarin Phillips Washington Brookings Inst., 1975

Zellner, A "An Efficient Method for Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias " *J American Statis Assoc* 57 (June 1962). 585–612

A Strategic Analysis of Monetary Interdependence

Koichi Hamada

University of Tokyo

Applying the monetary approach to the balance of payments, this paper examines the interdependent nature of monetary policies. In an *n*-country game where each monetary authority decides on its credit expansion to maximize its objective function, it is shown that the Cournot (or noncooperative) solution and the Stackelberg leadership solution do not lie on the contract curve, unless the aggregated preference over the balance of payments is exactly matched by the creation of international reserves. If the creation of international reserves exceeds the aggregated preference, the monetary expansion takes on the nature of public bads, leading to international inflation.

I. Introduction

In a system of fixed exchange rates, national monetary policies are closely interdependent. The same amount of monetary expansion of a country can have different effects on the domestic as well as the world economy if monetary policies taken by other countries in the system are different. Even in the case where, as at present, a system of floating exchange rates is partially adopted, the problem of the interdependence of monetary policies occurs within a monetary union if the monetary integration has proceeded to such a degree that the exchange rates between the participating countries are fixed but the monetary authorities are not completely unified.

McKinnon (1974) has proposed a new tripartite agreement for maintaining stable exchange-rates relations among the United States,

---

I am much indebted to Professors Jacob Frenkel, Harry G. Johnson, Jorge Marquez, and other members of the Trade Workshop at the University of Chicago, where a preliminary version of this paper was presented. Thanks are also due to Professors Hiroshi Atsumi, Michihiro Ohyama, Yusuke Onitsuka, and anonymous referees.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved.

Germany, and Japan. In order for such a proposal to be effective, we have to know the strategic positions of the countries participating in this monetary union within which exchange rates are more or less fixed. That is, we have to know what kind of incentives countries have when they decide on their monetary policies.

Most traditional approaches do not seem to pay due attention to the interdependent nature of monetary policies, since they either assume the monetary policies of other countries as given or take the international cooperation of monetary policies for granted.

The monetary approach to the balance of payments suggests that a disequilibrium in the balance of payments can be understood as a monetary phenomenon A surplus in the balance of payments appears when the demand for money of a nation exceeds the domestic supply of money, a deficit results when the domestic supply of money exceeds the demand for money At the same time, this approach enables us to investigate formally the structure of the interdependence of monetary policies Johnson (1972) derived a set of instructive formulae relating the rate of inflation, the balance of payments, the rate of growth of international money, and the rates of credit expansion of the various nations of the world One can see in these formulae the two sources of interdependence· (1) the world rate of inflation is the weighted average of domestic credit expansion, and (2) one country's surplus is another country's deficit

The purpose of this paper is to study the interdependent and strategic nature of monetary policies by directly applying the monetary approach to the balance of payments I set up a multicountry game situation in which the monetary policy of each country is conducted in such a way as to maximize the objective function of its monetary authority Monetary policies are interdependent in the sense that the effect of a specific policy adopted by a country differs depending on monetary policies taken by other countries In this framework, I shall ask how the difference in national preferences concerning inflation and the balance of payments affects the realized outcome of the rate of world inflation [1] I shall also relate the result to the recent development of the theory of public goods in political science (e g , Olson 1965)

Mundell (1972, chap 16) addresses the question of the relationship between world inflation and the international balance of payments using his policy assignment approach According to his analysis, the reserve-currency country should be assigned the objective of price stability and the other countries the objective of equilibrating their balances of payments In effect, this paper analyzes the same issue using an optimizing rather than an assignment approach (see Niehans 1968).

---

[1] For an analysis of the strategic implications of the interplay of national monetary policies in the framework of two-person game theory, see Hamada (1974a), where the monetary approach is partially adopted

In Section II, after presenting a simplified formulation of the monetary approach to the balance of payments, I shall construct a model of the game situation where the strategy of each country is the increase of its domestic credit. National monetary policies are assumed to be guided by the benefit-cost calculation of the monetary authorities. It will be shown that the Cournot solution (i.e., the intersection of the reaction curves) does not generally lie on the Paretian surface (i.e., the set of the cooperative policy configuration) unless the aggregated national preferences regarding the desired balance of payments of each country are matched by the creation of international reserves. If the creation of international reserves falls short of the aggregated preferences concerning balances of payments, then the monetary expansion takes on the nature of an increase in public goods and the Cournot solution tends to lie on the deflationary side of the Pareto-optimal configuration; if the creation of international reserves exceeds the aggregate preferences concerning balances of payments, then the resulting monetary expansion has the nature of public bads and the Cournot solution tends to lie on the inflationary side One can also obtain the Stackelberg leadership solution in the *n*-country case, provided that only one country behaves as the leader and the others as followers. The Stackelberg leadership solution also lies on the same side of the Paretian configuration as the Cournot solution

In Section III, the dependence of the final outcome on the preferences of monetary authorities will be illustrated by geometrical analyses, and derived explicitly for the case of a simple type of utility function I shall deal with the long-run implications of the analysis and the role of the reserve currency country as the leader country in Section IV In the concluding section, the application of our strategic analysis to the designing of a fixed exchange rate system, as well as its implications with regard to the system of floating exchange rates, will be discussed.

II. The Interplay of National Monetary Policies

Let me posit a world economy, or its subsystem, that consists of several countries among which the exchange rates are fixed Assume that goods are mobile enough for the assumption of the common price level to be justified [2] I shall concentrate my attention on the interplay of monetary policies I shall also assume that the utility function of a national monetary authority depends on the rate of inflation and the increase of its international reserves. More specifically, a monetary authority is assumed to have a most desired rate of inflation (or deflation) and a most desired value of increase (or decrease) in foreign reserves. Usually, the less the rate of inflation, the better will be the national welfare. But if the rate of inflation reduces below some specific value, further reduction of inflation, or further increase in the rate of deflation, may not be desirable.[3] Thus, I shall assume that there is a most desired rate of inflation or deflation for a nation that may differ from country to country. Similarly, a surplus in the balance of payments beyond a certain level may mean the reduction of consumption-investment opportunity for a nation. Thus, I shall assume that there is a most preferred value for the increase or decrease in international reserves that may differ from country to country

---

[2] It may be possible to extend our results to a more realistic situation where a national price level can deviate from the world price level, by appealing to the traded–nontraded-good model developed by Dornbusch (1973*b*).

As is well known, the discussion of the demand for international reserves is mostly concerned with the stock of reserves rather than the flow At any point in time, however, a movement to a desired level of stock can be expressed as a desired rate of change I shall interpret the most desired rate of the change in international reserves in this sense for the time being The relationship between the desired flow and the desired stock, that is, the relationship between the short-run and the long-run analysis, will be discussed thoroughly in Section IV

Suppose there are n countries in the system The exchange rates between these n currencies are assumed to be fixed as unity Let D_i, R_i, and M_i denote, respectively, the money supply due to credit expansion by the monetary authority, international reserves, and the demand for money of the ith country Then we have

$$D_i + R_i = M_i, \qquad i = 1, \quad , n \tag{1}$$

Here R_i is composed of gold, special drawing rights (SDRs), and reserve currency such as dollars[4] The real demand for money is assumed to be a function of national income Q and the money rate of interest r.

$$\frac{M_i}{p} = L^i(Q_i, r_i), \tag{2}$$

where p is the common price level in the world, $\partial L^i/\partial Q_i > 0$, and $\partial L^i/\partial r_i \leq 0$.

In order to focus our attention on the strategic structure of the interplay of monetary policies, I shall assume, for simplicity, that the effect of changing interest rates on the demand for money can be neglected The implications of relaxing this assumption will be discussed at the end of this section. Then, if we differentiate (2) with respect to time, noting that $\partial L^i/\partial r_i = 0$ and dividing by M_i, we have

---

[3] For discussions of the optimal rate of inflation, see, e.g , Bailey (1956), Phelps (1965), and Friedman (1969). In the actual world, wage and price rigidity would make further reduction in the rate of inflation below a certain level undesirable for a monetary authority. Such situations may not be fully consistent with our analysis below the point where the real rates of growth of nations are exogenously given. However, the strategic structure of the interplay of monetary policies would remain almost the same under such situations. See n 5 below.

[4] For the reserve country, we do not exclude the possibility that R_i might become negative. For example, if the amount of dollars held by foreign monetary authorities exceeds the reserve assets held by the United States, then R_i for the United States will be negative. We assume for simplicity that private sectors do not hold any amount of foreign currency, even though allowing this possibility does not change our results below.

$$\frac{\dot{D}_i}{M_i} + \frac{\dot{R}_i}{M_i} = \frac{\dot{p}}{p} + \eta_i \frac{\dot{Q}_i}{Q_i}, \qquad i = 1, \ldots, n, \tag{3}$$

where η_i denotes the income elasticity of the demand for money, i.e.,

$$\eta_i \equiv \frac{Q_i}{L^i}\frac{\partial L^i}{\partial Q_i}.$$

Define the ratio of the credit expansion to the demand for money in excess of the product of the growth rate of real national income and the income elasticity as x_i, and call it the rate of excess monetary expansion. That is,

$$x_i \equiv \frac{\dot{D}_i}{M_i} - \eta_i \frac{\dot{Q}_i}{Q_i} \tag{4}$$

Here the growth rate of real national income is assumed to be given exogenously,[5] and x_i corresponds to the strategy of the ith country in the terminology of game theory.[6] Also define the balance of payments of country i normalized by its demand for money as z_i, that is,

$$z_i \equiv \frac{\dot{R}_i}{M_i} \tag{5}$$

Then, by virtue of (4) and (5), (3) is reduced to, by denoting the common rate of inflation $\dot{p}/p$ as π,

$$z_i = \pi - x_i, \qquad i = 1, \ldots, n \tag{6}$$

Let the total amount of credit money, international reserves, and money demanded existing in the system be D, R, and M, respectively. Also let w_i be the relative share of the money demanded by country i, that is,

$$\sum_{i=1}^{n} D_i \equiv D, \qquad \sum_{i=1}^{n} R_i \equiv R, \qquad \sum_{i=1}^{n} M_i \equiv M,$$

and

$$w_i \equiv M_i/M, \qquad \sum_{i=1}^{n} w_i = 1, \qquad w_i > 0$$

---

[5] The endogenous treatment of the real rate of growth could be conducted along the following line. The recent theory of the Phillips curve tells us that the level of Q_i is related to the divergence of actual from expected inflation. Suppose that at a given moment the expected rate of inflation is fixed, then the level of output can be written $Q_i = f(\pi)$. However, since this complicates the formulation considerably, we will not go into the details of it here.

[6] We do not need to deal explicitly with the sterilization question, for our strategic variable D/M is the monetary expansion *in addition* to the monetary expansion due to a surplus in the balance of payments. If a country sterilizes surplus, then we can just treat D/M as negative.

Define the ratio of the increase in international reserves to the total world money as G_R such that $G_R = \dot{R}/M$.

By the definition of y_i and w_i, we have

$$\sum_{i=1}^{n} w_i z_i = \sum_{i=1}^{n} \frac{M_i}{M} \frac{\dot{R}_i}{M_i} = \frac{\sum \dot{R}_i}{M} = \frac{\dot{R}}{M} \equiv G_R. \tag{7}$$

Therefore, multiplying (6) by w_i and summing with respect to i, we obtain (noting $\sum_{i=1}^{n} w_i = 1$)

$$\sum_{i=1}^{n} w_i z_i = \pi - \sum_{i=1}^{n} w_i x_i. \tag{8}$$

Finally, we have from (7) and (6) the following two basic equations that are simplified representations of the monetary approach to the balance of payments.

$$\pi = \sum_{i=1}^{n} w_i x_i + G_R, \tag{9}$$

$$z_i = \pi - x_i \qquad \text{for } i = 1, \quad , n \tag{10}$$

Equation (9) indicates that the rate of inflation in the world economy depends on the weighted average of the excess supply of domestic credit expansion and the increase in international reserves. Equation (10) indicates that the balance of payments of a country depends on the difference between the weighted average of excess supplies of domestic credit of all the countries and its own excess supply of domestic credit These two equations are essentially the same as the equations obtained by Johnson,[7] with some simplifying notations here

Now let us turn to the objective functions of monetary authorities The objective function of each monetary authority is assumed to depend on the common rate of price increase and on the balance of payments of the country Let the utility function of the ith monetary authority be denoted by $u^i(\pi, z_i)$, where π is the rate of inflation common to all the countries and z_i is the ratio of increase in international reserves to the money demanded of the ith country.

As explained above, $u^i(\pi, z_i)$ is assumed to be a strictly concave function of π and z_i and to have the following properties

$$u_1^i = \frac{\partial u^i}{\partial \pi} \{\gtreqless\} 0, \qquad \text{according as } \pi \{\lesseqgtr\} a_i, \tag{11}$$

$$u_2^i = \frac{\partial u^i}{\partial z_i} \{\gtreqless\} 0, \qquad \text{according as } z_i \{\lesseqgtr\} b_i. \tag{12}$$

---

[7] Johnson (1972), pp. 1565 ff. Here I follow Johnson in that w_i's can be regarded as constants at the moment. For simplicity, we shall dwell on the continuous version of the monetary approach with smooth adjustment in the money market. For a different approach see, e.g., Dornbusch (1973*a*).

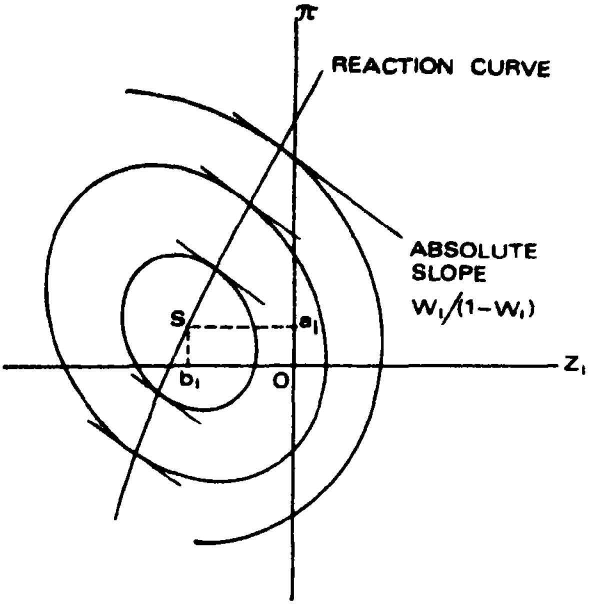


FIG 1 —Reaction curve with $a_i > 0$ and $b_i < 0$

Equations (11) and (12) indicate that country i has an ideal combination (bliss point) (a_i, b_i) of the rate of inflation and the rate of increase in international reserves The value a_i can be negative for a country that prefers deflation, b_i can be negative for a country that prefers a balance-of-payments deficit. Figure 1 indicates the indifference map for a country with a bliss point with $a_i > 0$ and $b_i < 0$ Indifference curves take circular shapes around the bliss point.

The game situation we have is thus the interplay of monetary policies x_i's by the monetary authorities The reaction curve for country i is obtained by maximizing $u^i(\pi, z_i)$ with respect to x_i, under the assumption that all other x_j's $(j \neq i)$ are given Noting that

$$u^i(\pi, z_i) = u^i\left(G_R + \sum_{k=1}^{n} w_k x_k, G_R + \sum_{k=1}^{n} w_k x_k - x_i\right),$$

we obtain

$$\frac{\partial u^i}{\partial x_i} = w_i u_1^i + (w_i - 1)u_2^i = 0, \qquad i = 1, \quad , n \tag{13}$$

The maximization condition (13) can be rewritten as

$$-\left.\frac{d\pi}{dz_i}\right|_{u^i \text{ const.}} = \frac{u_2^i}{u_1^i} = \frac{w_i}{1 - w_i} \tag{14}$$

The marginal rate of substitution between the rate of inflation and the growth rate of international reserves is equal to $w_i/(1 - w_i)$, which is larger for a country with a larger relative size in the world economy. The reaction curve is the locus of tangency points of the indifference curves to the lines with slope of $w_i/(1 - w_i)$. From figure 1, which indicates the reaction curve, one can see that the smaller a country is, the flatter the tangency slope, and that the reaction curve deviates more from the horizontal line $\pi = a_i$, indicating the preference of the ith country with regard to the rate of inflation. Smaller countries have a relatively stronger impact on their balance of payments than on the rate of inflation. The shape of the reaction curve is affected by the degree of this relative strength, which in turn depends on the size of the country. From (14), u_1^i and u_2^i have the same sign. Thus, $\pi > a_i$ implies $z_i > b_i$, and so forth. The country that desires a more stable price level tries to reduce the rate of inflation, sacrificing the target of balance of payments and incurring more surplus than is desirable. The exact point that a country actually chooses on the reaction curve depends on the behavior of the other countries. The Cournot solution is given by the intersection of the reaction curves (13).

On the other hand, the Pareto-optimal configuration for the world economy is given by maximizing

$$\sum_{k=1}^{n} \beta_k u^k(\pi, z_k)$$

with respect to x_i ($i = 1, \ldots, n$), where the parameter β_k is the relative importance, or the index of bargaining power, of country k, normalized as

$$\sum_{k=1}^{n} \beta_k = 1, \qquad \beta_k > 0 \tag{15}$$

The first-order condition for the Pareto-optimal configuration is

$$\sum_{k=1}^{n} \beta_k u_1^k w_i + \sum_{k=1}^{n} \beta_k u_2^k w_i - \beta_i u_2^i = 0, \qquad i = 1, \ldots, n \tag{16}$$

The Pareto-optimal surface, on which a cooperative solution would lie, is obtained by varying the weight β_k subject to (15).[8]

Summing equations (9) with respect to i, we obtain, noting that $\sum_{i=1}^{n} w_i = 1$,

$$\sum_{k=1}^{n} \beta_k u_1^k + \sum_{k=1}^{n} \beta_k u_2^k - \sum_{i=1}^{n} \beta_i u_2^i = 0,$$

or

$$\sum_{k=1}^{n} \beta_k u_1^k = 0 \tag{17}$$

---

[8] Mathematically speaking, the Pareto-optimal configuration is obtained by varying β_k within the interior of the $(n - 1)$ dimensional simplex.

If the most desired rate of inflation for country k is attained, then $u_1^k = 0$ Thus, equation (17) indicates that the Pareto-optimal configuration is achieved if the weighted average of the marginal costs of inflation is equated to zero [9]

On the other hand, summing equations (13) with respect to i, we obtain

$$\sum_{i=1}^{n} w_i u_1^i = \sum_{i=1}^{n} (1 - w_i) u_2^i. \tag{18}$$

Comparing (18) with (17), we can see that the Cournot solution does not generally lie on the Paretian configuration When the right-hand side of (18) is positive, international reserves and credit expansion have the nature of public goods, giving a deflationary bias to the Cournot solution When it is negative, international reserves and credit expansion have the nature of public bads, giving an inflationary bias to the Cournot solution For, in the former case of "public goods," each country welcomes a higher rate of increase in total money supply, but each country tries to expand its own money supply at a slower rate than others In the latter case of "public bads," each country welcomes a lower rate of increase in total money supply, but it tries to expand its own money supply at a higher rate than others

Suppose the most desired rate of inflation is identical for every country, that is, $a_i = a$ for any i (This includes the case where every country wants perfect price stability, i e , $a_i = 0$ for any i.) Then the Paretian configuration is naturally $\pi = a$ Since π is common to every country, u_1^i will have the same sign for every i, and accordingly u_2^i will have the same sign for every i More precisely, if the rate of inflation is larger than the optimal ($\pi > a$), then every country is getting more increase in reserves than is desired ($z_i - b_i > 0$), if the rate of inflation is lower than the optimal ($\pi < a$), then every country is getting less increase in reserves than is desired ($z_i - b_i < 0$) Take the case where $\pi > a$, then

$$\sum_{i=1}^{n} w_i z_i - \sum_{i=1}^{n} w_i b_i > 0 \tag{19}$$

By equation (7),

$$G_R - \sum_{i=1}^{n} w_i b_i > 0 \tag{20}$$

It is easy to see that equation (20) implies $\pi > a$

---

[9] In this particular sense, the Pareto-optimal configuration reflects the aggregation of the preferences of monetary authorities In fact, in the special case of quadratic utility functions analyzed in the next section, the Pareto-optimal rate of inflation is the weighted average of the desired rates of inflation, the weight being the relative importance of a country.

Thus, we can summarize: *Suppose the most desired rate of inflation is identical for each country. Then the Cournot solution gives a rate of inflation higher (lower) than the most desired rate of inflation if and only if the relative increase in international reserves is larger (smaller) than the weighted average of desired increases in international reserves.* In the situation with excessive expansion of world reserves, central banks defend themselves against reserve accumulation by expansion in domestic credit, thereby increasing world inflation (cf. Mundell 1972).

Under the same assumption of identical a_i, let us consider the effect of increasing the number of countries. Suppose the nth country is split into two identical countries with identical preferences on the rate of inflation and the relative increase in reserves, so that $\tilde{w}_n = \tilde{w}_{n+1} = w_n/2$ (new weights being denoted by $\tilde{w}_n$ and $\tilde{w}_{n+1}$). Then, as indicated by (14), the reaction curves of these two split countries deviate more from the Paretian line. Since the reaction curves of the first $(n - 1)$ countries remain the same, the Cournot solution will diverge more from the most desired rate of inflation. In fact, the same inflation rate and the same relative balance of payments $(\pi, z_1, \ldots, \tilde{z}_n, \tilde{z}_{n+1}$, where $\tilde{z}_n = \tilde{z}_{n+1} = z_n)$ will keep the left-hand side of (18) the same, but the absolute value of the right-hand side of (18) will be increased by the introduction of a new weight, $\tilde{w}_n = \tilde{w}_{n+1}$ $(= w_n/2)$. The equality of (18) can be restored only by increasing the divergence of π from the most desired rate of inflation a. Thus, splitting a country into two will make the Cournot solution diverge more from the Paretian configuration.

The Stackelberg leadership solution can be defined in this n-country case as long as a single country, say country 1, behaves as the leader and the others as followers. The leader maximizes its own satisfaction level $u^1(\pi, z_1)$ subject to the reaction curves of the others

$$w_i u^i + (w_i - 1)u_2^i = 0, \qquad i = 2, \ldots, n.$$

The leadership solution also does not generally lie on the Paretian configuration. Whether the Cournot solution or the leadership solution deviates more from the Paretian surface cannot be determined unambiguously, as is shown by the example in Section III. This leadership solution is of special interest in the dollar-standard system, where it is quite natural to regard the reserve country, the United States, as behaving also as the leader country. Since this situation is more relevant in the long-run context, we shall deal with the implication of the leadership solution in Section IV.

So far, we have neglected the effect of interest rates on the demand for money. If the rate of interest is changing, and if the interest elasticity for the demand for money is not negligible, then our analysis will have to be modified To illustrate this, let us take the particular money demand function

$$\frac{M_i}{p_i} = e^{-\lambda r} Q^{\eta_i}_i,$$

which yields

$$\frac{\dot{M}_i}{M} = \pi - \lambda r + \eta_i \frac{Q_i}{Q_i}$$

Suppose the real rate of interest can be regarded as constant, being eventually determined by the time preference of the public Then

$$\frac{M_i}{M_i} = \pi - \lambda\pi^e + \eta_i \frac{\dot{Q}_i}{Q_i}, \tag{21}$$

where π^e is the expected rate of inflation With adaptive expectation $\pi^e = \gamma(\pi - \pi^e)$,

$$\frac{M_i}{M_i} = (1 - \lambda\gamma) + \lambda\gamma\pi^e + \eta_i \frac{Q_i}{Q_i} \tag{22}$$

From equations (21) and (22), we can estimate the direction of the effect of the interest rate on the demand function of money When the rate of inflation is increasing, the expected rate of inflation tends to rise, thus reducing the demand for money This effect would make the reaction curve diverge more from the Paretian configuration to the inflationary side [10] On the other hand, when the rate of inflation is declining, the expectation for price decline is likely to occur, and this increases the demand for money This effect would let the reaction curve diverge more from the Paretian configuration to the deflationary side Therefore, it is generally expected that the effect of the interest rate would aggravate the divergence of the noncooperative solution from the cooperative solution but that the structure of conflict would remain unchanged Accordingly, in the next two sections we shall again neglect the effect of the changing interest rate

III. Illustrative Examples: Geometrical and Algebraic Analysis

In order to facilitate the understanding of the strategic structure of the monetary interplay of countries that may have different sizes and different preferences, let me present a geometrical analysis of simple cases At the same time, I shall derive various solutions under a specific form of the utility functions of monetary authorities as a guide to the geometrical analysis. The form of the utility function is

---

[10] With perfect foresight, π^e is equated with π Therefore, $\dot{\pi}^e$ becomes zero as far as the adaptive expectation formation is concerned, so that the modification in the form of (21) is necessary only when the expected rate lags behind the actual rate of inflation.

$$u^i(\pi, z_i) = -[(\pi - a_i)^2 + c(z_i - b_i)^2], \qquad i = 1, . . , n, \qquad (23)$$

where c represents the relative weight of the balance-of-payments consideration as compared with the rate of inflation. The following analysis is valid even with different c_i's for different countries, with some appearance of complication only in the formulae This form of the utility function gives indifference curves that are ellipses with axes parallel to the (π, z_i) axes. This simplification of the form of the utility function enables us to derive various solutions explicitly dependent on parameters such as a_i's, b_i's, w_i's, and c

Now let us start our geometrical analysis from a two-country case $(n = 2)$ where the two countries are of equal size $(w_1 = w_2 = \frac{1}{2})$. Suppose both countries want price stability $(a_1 = a_2 = 0)$ and a balance-of-payments deficit $(b_1 < 0, b_2 < 0)$ Also suppose the increase in international reserves is zero $(G_R = 0)$ Then, from equations (9) and (10) expressing the monetary approach, we obtain

$$\pi = \tfrac{1}{2}(x_1 + x_2), z_1 = \tfrac{1}{2}(x_2 - x_1), \text{ and } z_2 = \tfrac{1}{2}(x_1 - x_2)$$

In this simplest case, the rate of inflation is the average of the rates of excess monetary expansion, the balance of payments is proportional to the difference of these rates

The Stackelberg diagram is drawn like figure 2, the rates of excess monetary expansion x_1 and x_2 are the two coordinates The line $x_1 = x_2$ indicates the combination of monetary policies which would maintain the equilibrium in the balance of payments, $x_1 + x_2 = 0$ indicates their combination for price stability The indifference curves of country 1 are drawn with solid curves, which are oblique mirror images of indifference curves in the (z_1, π) plane (see fig 1), the indifference curves of country 2 are drawn with dotted curves, which are oblique transpositions of those in the (z_2, π) plane In other words, the π axis in the (z_i, π) plane in figure 1 coincides with the line $x_1 = x_2$, and the z_i axis coincides with the line $x_1 + x_2 = 0$ with inverse direction for country 1 and with the same direction for country 2 [11] The reaction curve of country 1 is the locus of the points where the tangency slope to the indifference curve of country 1 is horizontal, namely, the best points that country 1 can choose, given the rate of excess monetary expansion by country 2, that of country 2, the locus of the points where the tangency slope is vertical, namely, the best points that country 2 can choose, given the rate of excess monetary expansion by country 2 The Cournot solution N is the

---

[11] The slope of the indifference curves along the $\pi = 0$ locus is +1, and along the 45° line through S_1 it is −1

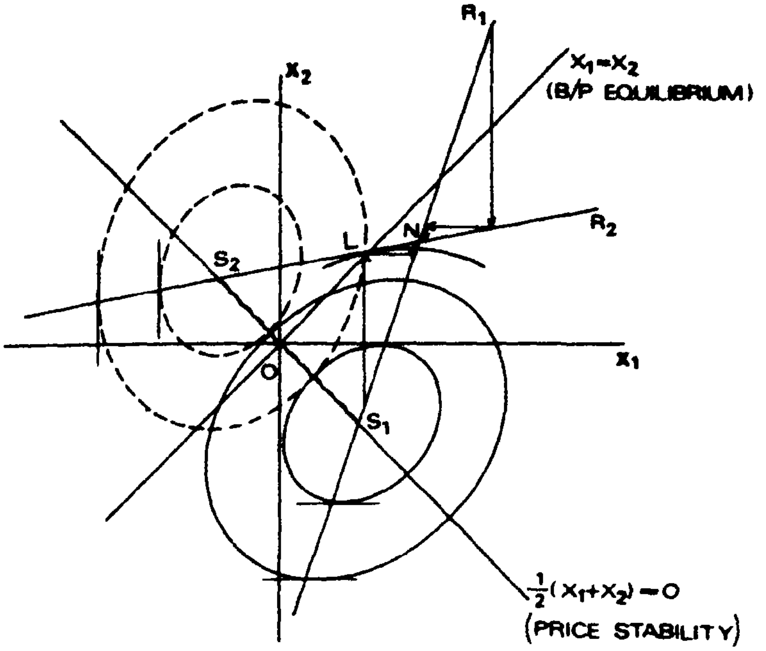


FIG. 2.—The case with $a_1 = a_2 = G_R = 0$, $b_1 < 0$, $b_2 < 0$, R_1, reaction curve of country 1, R_2, reaction curve of country 2, N, Cournot solution, L, leadership solution with country 1 as the leader, S_1S_2, Pareto efficient configuration

intersection of the two reaction curves It is a stable solution, as the arrows in the figure indicate. The leadership solution L_1 with country 1 as the leader is the point where country 1's indifference curve is tangent to the reaction curve of country 2, that is, the best point for country 1 on the reaction curve of country 2, the leadership solution with country 2 as the leader is not depicted but can be defined in a symmetrical way. The Pareto-efficient frontier, or the contract curve, is the locus of the points of tangency between the two indifference curves and the line segment on $w_1x_1 + w_2x_2 = 0$ (or $x_1 + x_2 = 0$ in this case) between S_1 and S_2.

From figure 2, representing the simplest case, we can note that neither the Cournot solution nor the leadership solution lies on the Paretian configuration (contract curve) and that both of them lie on the inflationary side of the Paretian configuration when b_1 and b_2 are negative. It is easy to see by constructing similar diagrams that, if both b_1 and b_2 are positive, then the Cournot solution as well as the leadership solution lies on the deflationary side of the contract curve If $b_1 + b_2 = 0$, then points S_1 and S_2 coincide, the contract curve shrinks to the same point, and the Cournot solution and the leadership solution coincide with the same

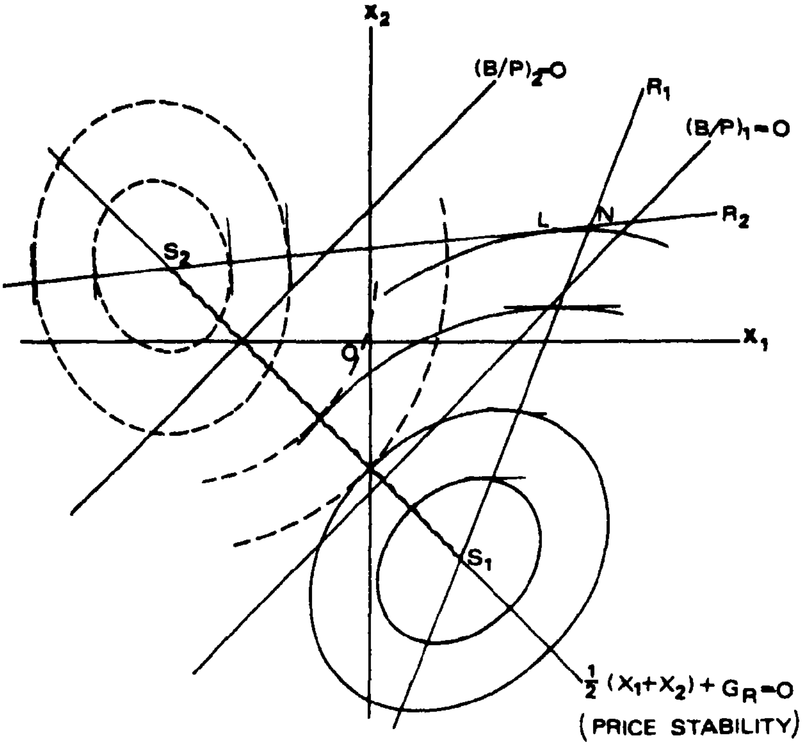


FIG 3 —The case with $G_R > 0$, $w_1 = w_2 = \frac{1}{2}$

point—the Paretian point We can do similar exercises varying the values of a_1 and a_2

Let us return to the case depicted in figure 2 where b_1 and b_2 are negative and $a_1 = a_2 = 0$. Suppose, however, there is some increase in international reserves That is, suppose $G_R > 0$. Then the price stability line becomes $\frac{1}{2}(x_1 + x_2) + G_R = 0$, shifting itself in a parallel way to the southwest direction The locus showing the equilibrium in the balance of payments for country 1 becomes $\frac{1}{2}(x_2 - x_1) + G_R = 0$, and that for country 2 becomes $\frac{1}{2}(x_1 - x_2) + G_R = 0$, shifting the equilibrium line of country 1 parallel to the southeast direction and that of country 2 to the northwest. Accordingly, the indifference map of country 1 moves to the south and that of country 2 to the west. Thus, we obtain the Stackelberg diagram as in figure 3. It can be seen that the divergence of the best points S_1 and S_2 is wider and that the Cournot as well as the leadership solution lies farther from the contract curve as compared with figure 2. However, in the case where both b_1 and b_2 are positive, the increase in G_R will contract the distance between S_1 and S_2, making the Cournot as

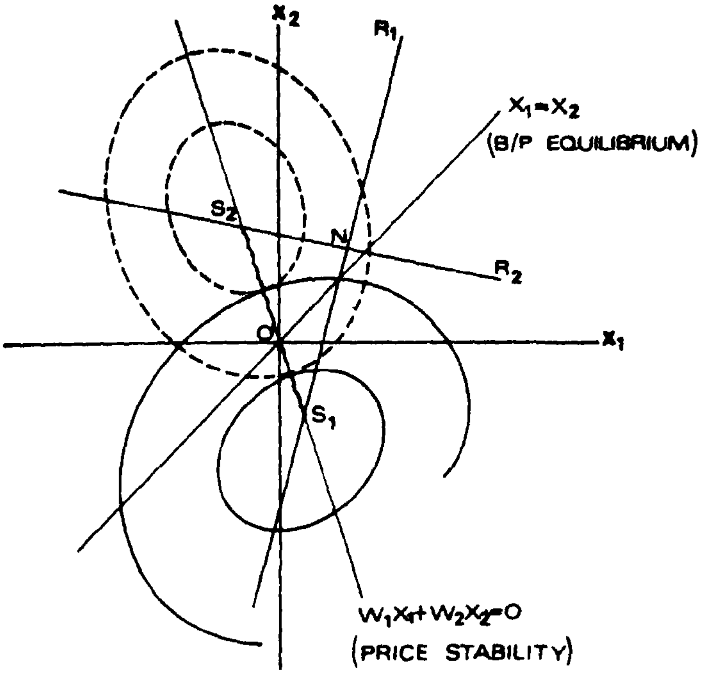


FIG 4 —The case with $w_1 > \frac{1}{2} > w_2$, $G_R = 0$

well as the leadership solution closer to the contract curve. In this case, the increase in international reserve assets contributes to improving the property of the noncooperative solution

Next we shall examine the effect of the relative size of a country. Starting again from figure 2 with $a_1 = a_2 = 0$, $b_1 < 0$, $b_2 < 0$, and $G_R = 0$, what happens to the diagram if w_1 is larger than half, that is, $w_1 > \frac{1}{2}$ and $w_2 < \frac{1}{2}$? The line indicating the balance-of-payments equilibrium remains the same, but the price stability line becomes steeper. Indifference curves will be skewed accordingly. From the discussion interpreting equation (14) in the last section, it can be seen that the reaction curve of the smaller country, that is, country 2, deviates more from the Paretian configuration, as is shown in figure 4. Therefore, the smaller country could exploit the larger country (see Olson [1965] and Olson and Zeckhauser [1966]), as far as the Cournot solution is concerned.

The understanding of the above geometrical analysis can be aided by considering the specific utility function (23) for the case where $n = 2$. The reaction curves are obtained by solving $\partial u^i/\partial x_i = 0$, or

$$w_i(\pi - a_i) + c(w_i - 1)(x_i - b_i) = 0, \qquad i = 1, 2. \tag{24}$$

The Cournot solution is the intersection of the reaction curves. The values of π, z_1, and z_2 can be obtained by solving a system of simultaneous equations (24) and (7), that is,

$$\sum_{i=1}^{2} w_i z_i = G_R. \tag{7}$$

Define a new weight $\bar{w}_i$[12] for country i as

$$\bar{w}_i \equiv \frac{(w_i^2)/(1 - w_i)}{\sum_{i=1}^{n} [(w_i^2)/(1 - w_i)]}, \qquad \bar{w} \equiv \sum_{i=1}^{n} \left(\frac{w_i^2}{1 - w_i}\right),$$

and the following weighted average

$$\bar{a} \equiv \sum_{i=1}^{n} \bar{w}_i a_i, \qquad \bar{b} \equiv \sum_{i=1}^{n} w_i b_i.$$

Then we obtain the formulae for the Cournot solution as

$$\pi = \bar{a} + \frac{c}{\bar{w}} (G_R - \bar{b}) \tag{25}$$

and

$$z_i = b_i + \frac{w_i}{c(1 - w_i)} (\pi - a_i), \qquad i = 1, 2 \tag{26}$$

From (25) we know that π equals the weighted average of a_i, the weight being the new weight $\bar{w}_i$, provided that G_R equals the simple weighted average of b_i. If the increase in international reserve assets falls short of the desired increase in international reserves by the monetary authorities, then there exists deflationary bias in the Cournot solution, if the increase in international reserves exceeds its demand, there exists inflationary bias in the Cournot solution.

The Pareto-optimal configuration, the cooperative solution, is given by maximizing

$$-\sum_{k=1}^{2} \beta_k[(\pi - a_k)^2 + (z_k - b_k)^2], \qquad \left(\sum_{k=1}^{2} \beta_k = 1, \beta_k > 0\right),$$

with respect to x_i ($i = 1, 2$) We obtain

$$-\sum_{k=1}^{2} \beta_k\{(\pi - a_k)w_i + c(\pi - x_k - b_k)(w_i - \delta_{ki})\} = 0, \qquad i = 1, 2, \tag{27}$$

where δ_{ki} is Kronecker's delta such that $\delta_{ki} = 1$ if $k = i$ and $\delta_{ki} = 0$ if $k \neq i$.

---

[12] This new weight $\bar{w}_i$ is monotonically related to w_i, the larger country being represented more than proportionately in the new weight. It is worth noting that $\bar{w}$ becomes smaller if a country is divided into smaller countries in the general case of $n > 2$.

Summing (27) with respect to i (cf. eq. [17]) gives

$$\sum_{k=1}^{2} \beta_k(\pi - a_k) = 0,$$

that is,

$$\pi = \sum_{k=1}^{2} \beta_k a_k = \hat{a} \tag{28}$$

On the Paretian frontier, π *should be equal to the weighted average of* a_k, weight β_k being the relative importance, or the index of bargaining power, of country k. Needless to say, $\bar{a}$ does not coincide with $\hat{a}$ unless all a_i are identical or $\beta_k = \bar{w}_i$ for all i. Comparing (25) with (28), we know that the Cournot solution does not in general lie on the Paretian frontier.

Similarly, we can derive the explicit formula for the leadership solution with country 1 as the leader and country 2 as the follower. For the follower, equation (24) is valid for $i = 2$.

$$w_2(\pi - a_2) + c(w_2 - 1)(z_2 - b_2) = 0 \tag{24'}$$

On the other hand, from (7) we obtain

$$w_2 z_2 = G_R - w_1 z_1,$$

so that (24′) is reduced to

$$w^*(\pi - a_2) - c(G_R - w_1 z_1 + w_2 b_2) = 0, \tag{29}$$

where

$$w^* = \frac{w_2^2}{1 - w_2} = \frac{w_2^2}{w_1}.$$

Country 1, the leader, maximizes

$$-[(\pi - a_1)^2 + c(z_1 - b_1)^2]$$

subject to the reaction curve (29) By equating the marginal rate of substitution of the isoquant of the objective function with that of (29) between π and z_1, we obtain

$$w_1(\pi - a_1) - w^*(z_1 - b_1) = 0 \tag{30}$$

The leadership solution is derived explicitly by solving (29) and (30) simultaneously.

$$\pi = \frac{1}{\Delta}[(cw_1^2 a_1 + w^{*2} a_2) + cw^*(G_R - b)], \tag{31}$$

$$z_1 = \frac{1}{\Delta}[w_1 w^*(a_2 - a_1) + cw_1(G_R - w_2 b_2) + w^{*2} b_1], \tag{32}$$

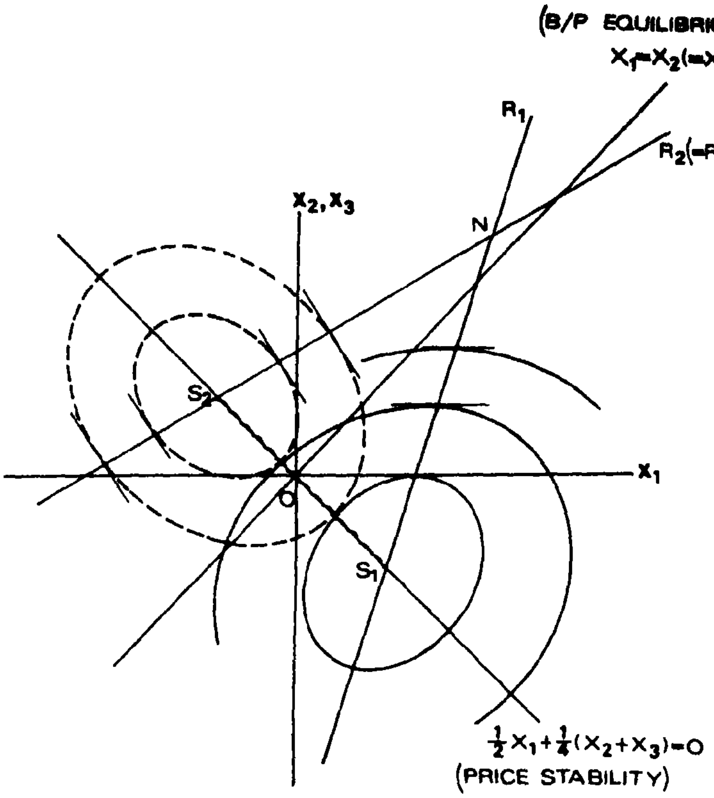


FIG 5.—The case with $w_1 = \frac{1}{2}$, $w_2 = w_3 = \frac{1}{4}$, $G_R = 0$

where $\Delta = cw_1^2 + w^{*2} > 0$ Whether the leadership solution deviates more from the Paretian frontier than the Cournot solution depends on the values of parameters, particularly the value of c, as is seen by comparing (31) with (25)

Finally, let us consider the case with more than two countries. Suppose the relative size of country 1 remains the same, that is, $w_1 = \frac{1}{2}$, but country 2 is divided into two smaller, identical countries with identical preferences, that is, $w_2 = w_3 = \frac{1}{4}$. If these small countries behave passively on the reaction curves, then they will behave identically. The slope of tangency of the reaction curve for the smaller countries is given by equation (14):

$$u_2^{\prime i}/u_1^{\prime i} = \tfrac{1}{4}/(1 - \tfrac{1}{4}) = 1/3, \qquad i = 2, 3.$$

Thus, the slope of tangency becomes less steep and the reaction curves of the smaller countries deviate more from the Paretian line, as is shown in figure 5. This tendency becomes more acute if we divide country 2 into many smaller identical countries. As pointed out by Olson, the increase in the size of a group tends to work against the optimal supply of public goods.[13]

IV. The Long-Run Implications and the Role of the Reserve Country

So far, we have carried out our analysis on the assumption that the utility functions of the monetary authorities are exogenously given. In particular, the value b_i, the most desired rate of change in international reserves, is assumed to be given. However, if we consider the dynamic process of adjusting the actual level of international reserves to the optimal levels, b_i should be interpreted as a parameter, being dependent on the current level of the stock of reserves. To incorporate this dynamic aspect would require a more complex model with sophisticated analytical methods such as optimal control and the differential game I do not intend in this paper to carry out this rather difficult task Instead, I shall clarify the nature of the long-run situations where the b_i's can roughly be interpreted to be equal to the desired rate of increase in international reserves, and describe the process of the short-run adjustment of the b_i's and the strategic interplays to this long-run state

When we deal with the long-run situation, the position of a reserve-currency country, that is, the country whose currency is used as international reserves, becomes significant, since the reserve country is the only country in our formulation whose R_i can be negative In the short run, R_i can be negative for any country, and accordingly b_i can be negative for any country that regards its current reserve holding as excessive compared with the desired level of stock However, since R_i cannot be negative for the non-reserve-currency country, b_i for these countries cannot be negative in the long run

In the long-run situation where the desired stock of international reserves is realized through time, let us suppose that the monetary

---

[13] The algebraic analysis in this section can easily be extended to the general n-country case The formulae for the Cournot and the Paretian solution hold true only for the index-of-summation runs from 1 to n. The leadership solution is also obtained if all countries except country 1 behave as followers In that case, the (aggregated) reaction curve along which the leader country, country 1, chooses the most favorable point can be written as follows [instead of (29)]

$$\sum_{i=2}^{n} \frac{w_i^2}{1-w_i}(\pi - a_i) - c\sum_{i=2}^{n} w_i(z_i - b_i) = 0$$

or, noting (7),

$$(\bar{w}^*\pi - \bar{a}^*) + c(w_1 z_1 + b^* - G_R) = 0, \qquad (29')$$

where

$$w_i^* = \frac{w_i^2}{1-w_i}, \quad \bar{w}^* = \sum_{i=2}^{n} w_i^*; \qquad \bar{a}^* = \sum_{i=2}^{n} w_i^* a_i, \quad b^* = \sum_{i=2}^{n} w_i b_i.$$

authorities try to keep constant the ratio of international reserves to nominal national income. Then the monetary authorities of these countries want to keep[14]

$$\frac{\dot{R}_i}{R_i} = \pi + \frac{\dot{Q}_i}{Q_i},$$

or

$$\frac{\dot{R}_i}{M_i} = \frac{R_i}{M_i}\left(\pi + \frac{\dot{Q}_i}{Q_i}\right).$$

Therefore, if we denote the optimal ratio of reserves to nominal income as θ_i, then on the ideal path

$$\frac{\dot{R}_i}{M_i} = \theta_i\left(\pi + \frac{\dot{Q}_i}{Q_i}\right), \qquad \text{or} \qquad z_i - \theta_i\pi = \theta_i\frac{\dot{Q}_i}{Q_i}.$$

We can see that, under the given growth rate of real output, the monetary authorities of nonreserve countries want to keep the normalized *real* change in reserves as constant. Thus, in the long run, we can redefine the preference function of monetary authorities in terms of the rate of inflation π and the real change in reserves $(z_i - \theta_i\pi)$ such that $u^i(\pi, z_i - \theta_i\pi)$ Here b_i's can be regarded as positive constants equalizing $\theta_i\dot{Q}_i/Q_i$. Since $z_i = \pi - x_i$, the utility function can be rewritten

$$u^i[\pi, (1 - \theta_i)\pi - x_i]$$

Formally, the same thing applies to the reserve country, except that b_i and θ_i can be negative Because it is safe to assume θ_i is less than unity, the analysis in the previous sections holds without any essential change except the inclusion of θ_i For example, instead of (13), the reaction functions can be written

$$\frac{\partial u^i}{\partial x_i} = w_i u_1^i + [(1 - \theta_i)w_i - 1]u_2^i = 0$$

The essential nature of policy interplay remains unchanged in the presence of the compensating term for the capital loss (or gain) in international reserves[15]

Thus, we can almost regard the analysis in the previous sections as the description of the long-run situation. The only reservation is that only the reserve-currency country, say country 1, can have negative b_1 and that the other countries should have positive b_2 in the long run Therefore, the Stackelberg diagrams in the long run should be drawn with b_1 negative and other b_i's positive. The lesson we obtained in the previous sections is that if

---

[14] If the nonreserve countries want to keep constant the ratio of international reserves to the outstanding domestic money supply, then $\dot{R}_i/R_i = \pi + \eta_i\dot{Q}_i/Q_i$

[15] Even in the short-run analysis, capital loss may have to be taken into account if we make the perfect-foresight assumption But since b_1 itself is changing, we do not formally incorporate it into the model.

$$G_R - \sum_{i=1}^{n} w_i b_i > 0 \qquad (20)$$

then world inflation is likely to occur for the case of identical a_i's. Suppose country 1, the reserve country, has a negative b_1 and other countries positive b_i's. Then the condition becomes

$$G_R + w_1 |b_1| > \sum_{i=2}^{n} w_i b_i$$

Thus, the Cournot solution will lie on the inflationary side if the (aggregated) desired increase in international reserves by the non-reserve-currency countries is more than matched by the creation of international reserves in the form of international money and the supply of money of the reserve currency.

The adjustment process of the short-run to the long-run situation can be visualized in figure 2. Suppose, at the initial moment, country 2 has accumulated an excessive stock of reserves and tries to decumulate it. Then b_2 as well as b_1 is negative, leading to the Stackelberg diagram like figure 2. However, when in the process of adjusting, the actual stock of country 2 approaches the desired stock of reserves, then b_2 becomes positive. The position of the Cournot solution in the long-run situation depends on the relative magnitude of the absolute value of $w_1 b_1$ and $w_2 b_2$.

In the long run, it is quite likely that the only country that can behave as the leader is the reserve-currency country. Our analysis implies that, if the reserve-currency country keeps b_i negative with a large absolute value, then the noncorporative solution as well as the leadership solution tends to be inflationary. In this respect, the failure of the dollar-standard system can be regarded as substantially due to its particular setup of the rule of the game that only the United States can have a large deficit, that is, a negative b_i with a large absolute value, and that only she can behave as a leader. The system would have worked properly only if the United States behaved as a benevolent or paternalistic leader considering only the world rate of inflation, rather than as a leader in Stackelberg's sense.[16]

The example illustrated in figure 3 suggests the role of the international organization. If the international monetary organization sets its relative change G_R in international money to equate the (net) aggregate desired increase in international reserves, that is,

$$G_R = \sum_{i=1}^{n} w_i b_i,$$

---

[16] Mundell (1972, chap. 16). For a discussion of the two concepts of leadership, see Hamada (1974b).

then the noncooperative behaviors of national monetary authorities will achieve the Paretian configuration, because the Cournot, the leadership, and the Paretian solution collapse into a single point In the long run, G_R must be positive, so that the optimization through the international monetary organization is workable as long as $\sum_{i=1}^{n} w_i b_i$ is nonnegative. When the reserve-currency country desires so large an amount of deficit that

$$\sum_{i=1}^{n} w_i b_i < 0$$

or

$$-w_1 b_1 > \sum_{i=1}^{n} w_i b_i,$$

then the role of the international organization to induce the noncooperative solution to the Paretian solution is unworkable

In the terminology of game theory, the international organization, under the favorable conditions mentioned above, can change the payoff structure of our monetary game from a game akin to the prisoner's dilemma to the game in which the noncooperative solution coincides with the cooperative solution In fact, this is analogous to an approach in peace research in political science that seeks the method of changing the payoff structure of the prisoner's-dilemma case for the better

V. Concluding Remarks

Despite its simplified structure, I hope this analysis contributes to an understanding of the structure of monetary interdependence in the fixed-exchange-rates system or a subsystem of the world economy in which exchange rates are fixed among participants The examples in Section III reveal the way the final outcome depends on the aggregated preferences of monetary authorities over the rate of inflation and balances of payments

If mutual cooperation is possible, then the divergence of various solutions from the cooperative solution, that is, the Paretian configuration, is not so important. Countries will cooperate to achieve the outcome on the Paretian frontier. The only remaining issue, which can become politically serious, is how to achieve the point on the Paretian surface that is relatively favorable to a particular country—in our context, how to achieve the Paretian solution with a higher value of β_i for country i

Cooperation, however, is not always possible In this case, in order for any fixed-exchange-rates system to work successfully, the system must be designed in such a way that the noncooperative solution, that is, the result of individual rational behavior, will not locate itself very far from the cooperative solution. My analysis suggests that there are at least two ways of achieving this design One way is to manipulate the rate of increase in international reserve assets to match the average preference for accumulating or decumulating reserves by the monetary authorities Another way is to devise a clearing system so as to keep the preferences of monetary authorities for surplus or deficit to suitable values.

Finally, let us briefly consider the application of the above analysis to the flexible-exchange-rate system As was shown elsewhere, the conflicting situation is mostly eliminated in the system of flexible exchange rates In the short run, conflicts over the level of interest rates may occur, and in the adjustment process there may be phenomena like the beggar-my-neighbors policy (Hamada 1974*a*) But in the long run, where the real rate of interest is determined by the time preferences of the participating nations, even the conflict over the rates of interest is unlikely to occur

.The above analysis suggests, however, that one possible source of conflict lies in a wrong conception of national interests Suppose national monetary authorities base their welfare functions partly on the balance of the current account rather than the overall account Then some monetary authorities may sacrifice price stability for an increased surplus in the current account by competitive devaluation If such a situation arises, the preference for a surplus in the balance of the current account might aggravate world inflation in the system of flexible exchange rates, while in the system of fixed exchange rates, the preference for a deficit in the balance of the overall account might aggravate world inflation

References

Bailey, M J "The Welfare Cost of Inflationary Finance" *J P.E.* 64, no. 2 (April 1956) 93–110

Dornbusch, R. "Currency Depreciation, Hoarding, and Relative Prices." *J.P.E.* 81, no 4 (July/August 1973) 893–915 (*a*)

——— "Devaluation, Money, and Nontraded Goods." *A.E.R.* 63 (December 1973) 871–80. (*b*)

Friedman, M. *The Optimum Quantity of Money* Chicago Aldine, 1969.

Hamada, K "Alternative Exchange Rate Systems and the Interdependence of Monetary Policies" In *National Monetary Policies and the International Financial System*, edited by Robert Z. Aliber. Chicago Univ. Chicago Press, 1974. (*a*)

———. "On the Political Economy of Monetary Integration. A Public Economic Approach." Presented at the Wingspread Conference II, July 1974. (*b*) (To appear in *Political Economy of Monetary Reform*, edited by R. Z. Aliber. London: Macmillan)

Johnson, H. G. "The Monetary Approach to Balance-of-Payments Theory." *J. Financial and Quantitative Analysis* 7, no. 2 (March 1972)· 1555–72.

McKinnon, R. I. "A New Tripartite Agreement or a Limping Dollar Standard?" In *Essays in International Finance*. Princeton, N.J.. Princeton Univ. Press, 1974.

Mundell, R. A. *Monetary Theory: Inflation, Interest, and Growth in the World Economy.* Pacific Palisades, Calif.: Goodyear, 1972.

Niehans, J. "Monetary and Fiscal Policies in Open Economies under Fixed Exchange Rates. An Optimizing Approach." *J.P.E.* 76, no. 4, pt. 2 (July/August 1968): 893–943.

Olson, M., Jr. *The Logic of Collective Action: Public Goods and Theory of Groups.* Harvard Economic Studies, vol. 124. Cambridge, Mass.: Harvard Univ. Press, 1965.

Olson, M., and Zeckhauser, R. "An Economic Theory of Alliance." *Rev. Econ. and Statis.* 43 (August 1966): 266–79.

Phelps, E. S. "Anticipated Inflation and Economic Welfare." *J.P.E.* 73, no. 1 (February 1965): 1–13.

Migration, Age, and Education

Aba Schwartz

Tel-Aviv University

The shape of the earnings-age function and the way education affects it can explain all the empirically observed relations of migration measures to distance moved, age, and education. No other assumptions (which are often made)—such as attributing lower risk aversion and higher efficiency of job search to more-educated persons, or higher psychic cost of moving to older persons—are needed. The individual's search for jobs and the firm's search for employees are affected by the shape of the earnings function—and, thus, by age and education in a predictable way. In turn, the observed systematic behavior of the migration measures with respect to age and education is responsive to the search behavior

I. Introduction

The interrelationships between migration, age, education, distance moved, and earnings are discussed in numerous economic studies. One of the earliest is Sjaastad (1961, 1962) It has been established that, on the average, migrants move from lower-income regions to higher-income regions, and that migration flows (controlling for population base) increase with education and diminish with age and distance moved. Folger and Nam (1967), Shryock (1959), and Schwartz (1971) show that the ratio of net to gross migration declines with education.

There is considerable overlapping of interpretations of these findings in the economic literature, since many of the studies have their roots in human capital theory. However, there are conflicting interpretations of

This study was sponsored by the Center for Human Resource Research, Ohio State University, and the David Horowitz Institute for the Research of Developing Countries, Tel-Aviv University. I wish to thank G. Nestel for hours of penetrating discussions and suggestions. I am also indebted to V. Adams, T Hatta, J. Davis, and the Ohio State University Labor Economics Workshop.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved.

some observed phenomena—particularly with respect to the impact of distance. Stouffer (1940) attributed the sizable negative effect of distance to "intervening opportunities"—suggesting that, the longer the distance, the more opportunities are encountered by a searching individual.[1] Building on Stouffer's theory, Levy and Wadycki (1974*a*) introduced a set of distance-related variables they define as the best alternatives and construct as follows. For any origin and any radius of search from that origin, the best alternative variable (say, wage) is the largest value the variable receives anywhere within that radius. The introduction of these variables reduces (in a regression analysis) the coefficient of distance[2] (but in no way wipes it out), proving that distance is at least partially a proxy for alternative opportunities Brennan (1965) attributes the negative effect of distance to increasing psychic cost, Schwartz (1973) attributes it to diminishing information

Classifying the migration measures by education generates more confusion Courchene (1970) and Schwartz (1973) established that more-educated individuals move, on the average, longer distances. A few alternative hypotheses have been offered. The efficiency hypothesis suggests that the search-production function is technologically more efficient as education increases, that is, given the cost of search, more-educated individuals gain more information on faraway jobs The psychic cost hypothesis is a two-stage argument It suggests that psychic cost of migration increases with distance and that education reduces its importance. Between the two hypotheses, the efficiency hypothesis seems more convincing on the basis of an indirect test performed by Schwartz (1973).

Recently, David (1974) developed a comprehensive theory of migration as a personal decision-making problem, in which distance moved (migration) is but one dimension of a job search—the other being the number of searches. In conjunction with other fairly general assumptions, it develops some useful theorems which link the distance moved and the number of searches to initial wealth and to behavior toward risk His theoretical framework brings forward some more hypotheses when he attempts to use his theory to explain certain empirical regularities. The diminishing rate of migration with age can be obtained only by assuming that risk aversion increases with age (or, alternatively, that psychic cost increases with age) faster than does wealth Likewise, it must be assumed that more-educated persons are less risk averse than are less-educated persons, in order to obtain the empirically observed phenomena of more-educated persons moving, on the average, longer distances than do less-educated persons.

---

[1] Unless we assume that an individual starts wandering and stops only when he finds an appropriate job, the intervening opportunities theory is basically an information hypothesis. A test of the intervening opportunities hypothesis was conducted by Miller (1972).

[2] The diminishing impact of distance may be purely statistical, since by construction the best alternative variables are monotone functions of distance. Thus, if the original relation of migration to distance is not linear, the newly added variables will capture part of the distance effect.

Neither the efficiency hypothesis nor David's hypotheses have been directly tested. Moreover, as I show in Section V, the rate of migration diminishes faster with age as education increases. This result holds even after I account for economic differences among regions in the form of the present value of returns to migration by age, by education, and by regions. None of the discussed hypotheses can explain this systematic variation—which strikingly resembles the systematic variation of earnings with age and education (earnings increase faster with age as education increases) This similarity suggests that the variation in the rate of migration (by age and education) depends on the same human capital factors which determine the corresponding variation in earnings, we show that such a dependence does exist Moreover, other empirical variations of the migration measures can as well be attributed to the same widely accepted human capital arguments. This unified approach eliminates the need to resort to ad hoc assumptions for each of the observed empirical regularities

I cannot and do not claim that the efficiency assumption (or any of the other assumptions discussed earlier) is wrong. Rather, I claim that some of the empirical phenomena need human capital arguments to explain them and that the rest are easily explained by the same arguments—thereby generating a simpler theory (in the sense that fewer assumptions are needed to account for all the phenomena) Moreover, the human capital assumptions used in this paper are more acceptable on empirical grounds, as they are based on an intensive study of the earning function, whereas no direct study sheds light on the alternative hypotheses discussed here

In Section II we restate some human capital hypotheses and derive some consequences regarding the cost of job switching In particular we discuss how age and education affect this cost In Section III we discuss the firm's search for employees, the individual's search for jobs, and how the search is affected by age and education In Section IV we derive propositions on migration, age, and education and refer to some corresponding empirical findings In Section V we provide an empirical study which specifically relates age, education, the rate of migration, and distance moved We conclude in Section VI

II

Formal education provides general skills which are transferable; it also provides some skills which are specific to a given occupation. Furthermore, increased formal education increases the potential of learning over time through experience on the job. Put differently, certain types of jobs offer inherent opportunities for more and faster promotion as experience is gained over time—and the more formal education an individual has, the more such jobs are open to him. This on-the-job experience has specific components. The share of the specific components in the total experience does not decline as education increases, and it is lost if the individual switches occupations. These arguments are standard human capital theoretical explanations for the observed phenomena that the earnings function increases with age, that it is steeper with higher education, and that the earnings function of more-educated persons dominates the function of less-educated persons soon after they complete their formal education [3]

Assume (for simplicity of exposition) that in labor market equilibrium all occupations requiring the same amount of formal education have identical earnings functions for any age at which the individual starts that occupation. Hence, in equilibrium, the earnings function $Y(E, B, A)$ varies only with the education level E, the age at which the individual started that occupation B, and his current (or any future) age A Let $B(E)$ be the age of graduation when E is the level of required education. Obviously $B > B(E)$ implies that the individual switched to his current job (when he was B years of age)

Define the earnings-inferiority function of an occupational switcher as follows

$$L(E, B, A) = Y[E, B(E), A] - Y(E, B, A) \tag{1}$$

Our discussion in the first paragraph implies that.

$$\frac{\partial L}{\partial E} > 0, \frac{\partial L}{\partial B} > 0, \frac{\partial^2 L}{\partial E\, \partial B} > 0 \tag{2}$$

The present value of L is the opportunity cost of occupational mobility

We now turn to the direct cost of mobility, which may include geographical mobility as well The cost of geographical mobility is composed of elements which do not depend on the distance moved, such as the transaction costs of selling property in the place of origin and buying comparable property at the destination, and the distance-related transportation cost It seems plausible to assume that these cost items do not diminish as age or formal education increase [4] The direct cost of occupational mobility may involve payments for concentrated studies of the new skills and for the necessary materials and books. It seems safe to assume that these cost items do not diminish with age or education and that, if an interaction between age of switching and education exists, it is not negative

---

[3] See Becker (1964) for a comprehensive statement of the human capital theory.

[4] This assumption does not imply that mobility is costless.

Let $C(E, B)$ be the cost of mobility at age B, for a person with education level E. Our discussion implies:

$$\frac{\partial C}{\partial B} \geq 0; \frac{\partial C}{\partial E} \geq 0; \frac{\partial^2 C}{\partial B \partial E} \geq 0. \tag{3}$$

Finally, let ρ be the discount rate and T the retirement age. The cost of mobility at age B, if the economy is in equilibrium, is:

$$Z(E, B) = \int_B^T L(E, B, A) e^{-\rho(A-B)}\, dA + C(E, B) \tag{4}$$

From (4) we get the cost per unit of time $P(E, B)$ of an occupational switcher[5] at age B, whose education level is E:

$$P(E, B) = \frac{Z(E, B)}{\int_B^T e^{-\rho(A-B)}\, dA} = \frac{\rho}{1 - e^{\rho(T-B)}} Z(E, B). \tag{5}$$

From (2) and (3) we obtain:

$$\frac{\partial P}{\partial E} > 0, \frac{\partial P}{\partial B} > 0; \frac{\partial^2 P}{\partial E \partial B} > 0 \tag{6}$$

That is, in equilibrium, the per-unit time cost of switching increases with the required education and with the switching age. It has a positive interaction between the switching age and the required education

It was easier for the sake of exposition to start from equilibrium as we defined it in the text, as this freed us from the need to have indexes for all occupations However, the result expressed in (6) is quite general. We restate it as·

> Proposition 1. The per-unit time switching cost of a job-seeking individual increases with age at a rate which increases with education [6]

Proposition 1 considers the switching cost from the job-seeking individual's point of view The argument is symmetric for the firm's search for employees Assume that individuals won't switch unless their switching cost is covered by employers; in this case, the cost of switching is transferred to the employers—and has the same systematic behavior with respect to age and education of the employees. We summarize as:

---

[5] Switching an occupation may involve geographical mobility, as discussed above.

[6] Note that individuals of the same age and occupation differ among themselves in many respects; thus, their alternative costs of switching and earnings inferiority in the new job vary. Obviously, we may have two persons holding similar jobs while the alternative cost of switching is higher for the younger. Likewise, we may find that the alternative cost of a less-educated person exceeds that of a more-educated one. The arguments developed above relate to the representative person in each education-age group. Once the individuals' variation is taken into consideration, the proposition is correct in a statistical sense.

> Proposition 2: The cost to the firm of inducing potential employees to switch from other occupations to those needed by the firm increases with the age of the potential employee at a rate which increases with the required education.

Equipped with propositions 1 and 2, we now consider the search of a firm for employees and the search of an individual for a job.

III

Migration in a homogeneous country is primarily a search for jobs. In order to understand this phenomenon, we have to understand the economic forces underlying the search process—the forces working on the individual who looks for a job and those working on the firm that looks for employees

Our discussion presumes a homogeneous country in terms of the distribution of cultural and economic variables Thus, from the job searcher's point of view, jobs are generated randomly over the area The probability distribution of the number of jobs and their wage structure—in any occupation, throughout any given period of time—is the same anywhere In a likewise simplifying manner, to help us focus on the pure search behavior, we assume that each firm faces a homogeneous area with no region-specific differences in the distribution of manpower by specific skills or age The age-education-specific conclusions we derive, however, are general Unequal regions will differ only in the sizes of the flows of migrants—producing unequal flows into and out of different regions Some regions may be net recipients, whereas others may be net senders of migrants [7]

The Firm's Search for Employees

The exact market structure in which the firm operates is irrelevant to our analysis; it is not important whether the firm is competitive, or whether it is oligopolistic or monopolistic All we need for the validity of our analysis is that the firm minimizes cost when it attempts to attain its targets, that the firm's search process is not costless, and that it is characterized (like any other productive activity) by a search-production function. The search is an active one. The firm generates information on its needs by advertising in labor exchanges, placement offices, and other forms of labor markets and media pipelines. It also uses interviews to screen these marketplaces for available potential employees with the needed skills.

---

[7] Migration is not necessarily a response to some general measure of regional economic differences (in a homogeneous country such as the U.S.), but rather a response to personal opportunities, this is demonstrated by the fact that for any two regions we observe flows of migrants in both directions. In fact, the net flow is meager even in the presence of regional differences (Sjaastad 1962). Note also that the highest rate of migration is observed for the most educated-persons, for whom regional differences are nonexistent or very small (Schwartz 1971).

This search can be extended spatially to include exchanges and other channels which are farther away, thereby enabling the firm to screen a wider area (population base) if more recruits are needed. The firm can increase the number of recruits within a given area by increasing the share of the eligible population considered for the job; it does this by switching potential candidates to its specific occupational needs—starting with individuals having occupational characteristics closest to those required, and going all the way to the periphery of the population available for switching We can thus view the search procedure for any given occupational requirement as a two-dimensional production function Its output is the number of needed persons, its two inputs are *R*, the area of search, and *S*, the share of the eligible population considered for the job

Note that the technological efficiency of the firm's search for employees does not vary with the latter's required skills (and thus with their education), since any rational firm will always use the most efficient technology available to it However, economic (allocative) efficiency results in a different input mix in the search function for different types of skills In particular, the input mix varies with education

Whether a search is more *R*-intensive or *S*-intensive depends on the ratio between the corresponding marginal factor costs of the inputs. We focus on the way age and education affect this ratio

If we hold the marginal cost of switching inducement constant, then the higher the education required to perform a certain job, the smaller will be the share of the eligible population that can be induced to switch. This follows from proposition 2 (through eq. [6]). First, the cost of switching inducement increases directly with education Second, the marginal cost of switching inducement rises faster with age as the required education increases, which further decreases the share of those that can be induced to switch, given the marginal cost Put another way· Given the share of those that can be induced to switch (out of the eligible population), the marginal cost of switching inducement increases with the required education.

On the other hand, in order to increase the areas of search, the firm has to dispatch more recruiting services to more distant locations; the associated marginal factor cost is distance related and depends neither on the required education nor on the age of the prospective candidates.

For simplicity, we assume that the population eligible for switching into an occupation with a required education level E is all the population with education level E.[8] Let $\alpha(E)$ be the number of persons with education level E per unit area (the population density); $N(E)$, the number of needed employees of education level E; $C(R, N)$ the marginal cost per person searched (if N persons are needed) of adding one more unit of area R to the search (as mentioned earlier, it does not depend on E); and $P(S, E)$, the marginal cost per person searched of increasing by one unit the considered share (of the eligible population) for switching S. Let

$$N(E) = \alpha(E) f(S, R) \tag{7}$$

be the production function of recruits; it is identical for all education levels up to $\alpha(E)$ Factors S and R are the instruments of search (the "factors" of production), to be chosen optimally. Here $f(S, R)$ does not depend on E, since the firm uses its most efficient technique of search for all the education categories Thus, for any choice of S and R, the number of recruits obtained differs among education groups only by the corresponding population density factor, $\alpha(E)$. Our discussion of the firm's behavior suggests the following minimization problem as far as search is concerned

$$\operatorname*{Min}_{R,S} \left\{ N(E) \left[\int_0^R C(u, N)\, du + \int_0^S P(v, E)\, dv \right] - \lambda[\alpha(E) f(S, R) - N(E)] \right\}, \tag{8}$$

where λ is the Lagrangian multiplier

The first-order condition of this problem yields

$$-\frac{dR}{dS} = \frac{f_S}{f_R} = \frac{P(S, E)}{C(R, N)} \tag{9}$$

Differentiating (9) with respect to E yields

$$-\frac{\partial}{\partial E}\left(\frac{dR}{dS}\right) = \frac{P_E(S, E)}{C(R, N)} > 0 \tag{10}$$

The inequality in (10) follows from (6) and implies that R/S is an increasing function of E for any N[9] Thus, the search becomes more area intensive as the required education increases This does not imply that the actual area of search increases; the area also depends on the number of recruits searched for in each education category. If the number of recruits needed decreases with the required education, the area of search may diminish. However, note from (8) (after dividing it by $\alpha(E)$) that if:

---

[8] In reality, only a subgroup of the population with education level E is eligible for any particular job requiring that level of education If we take that into consideration, the formal presentation will be more cumbersome, since a job index would have to be included. However, the results derived from the simplified presentation will be sustained.

[9] This result requires $f(S, R)$ to be a quasiconcave function, with positive marginal productivities. This in turn requires that holding S constant and doubling the area of search will at least double the number of recruits (a safe assumption to make).

$$\frac{N(E')}{\alpha(E')} \geq \frac{N(E)}{\alpha(E)} \quad \text{for any } E' > E, \tag{11}$$

the area of search increases with the required education. Inequality (11) is logically plausible and is supported by observed migration flows for any geographical division[10] Equations (10) and (11) imply the following·

> Proposition 3: The firm's area of search for employees to perform a given job increases with the job's required education.

Equations (10) and (11) also imply that S (the share considered for switching) diminishes with the required education This is due partly to the diminishing (with education) set of occupations from which individuals can be induced to switch, given the marginal cost of switching; more important to our analysis, however, is that it occurs because the age range for switching diminishes with the required education

The firm does not require a specific age-education distribution of employees, rather, it needs a specific distribution of persons by their capability of performing the required variety of occupational tasks It is virtually impossible to substitute less-educated recruits for recruits with the education required by a given occupation On the other hand, the firm can substitute younger recruits for older ones provided they have the same (or very similar) educational backgrounds Proposition 2 then implies that the age range of the employees whom the firm will attempt to induce to switch diminishes with the required education When individual randomness is superimposed on earnings means (for age-education categories), proposition 2 implies that—starting with the age at which the most educated individuals graduate—the relative frequency distribution of recruits by age declines with age at a rate that increases (in absolute value) with education Put another way, the cumulative relative frequency distribution of more-educated individuals dominates the corresponding distribution of less-educated individuals. In particular, the relative frequency of the youngest age group (in the considered age range) increases with education, and the relative frequency of the oldest age group declines with education This leads to

> Proposition 4 Starting with the age at which the most-educated individuals graduate, the relative frequency distribution of recruits by age declines with age at a rate that increases (in absolute value) with education.

---

[10] See, e.g., Levy and Wadycki (1974a).

The Individual's Search for a Job

The discussion about the firm's recruiting behavior centered on the search-production function. Similarly, we may view each individual's search for a job as a trade-off between waiting time and the size of the area searched If the individual's strategy is to accept the first job which meets (or exceeds) his reservation wage,[11] his waiting time until he finds one diminishes if he searches a larger area. Of course, screening a larger area involves a higher search cost, which depends on neither age nor education For our purpose, we have to establish how age and education affect the waiting time, given the area of search.

The economic feasibility of switching requires that there be positive gains from it By proposition 1 (and equation [6]), the economic feasibility of occupational switching diminishes with age and education; this, in turn, limits (within a given area of search) the number of job openings which exceed the individual's reservation price. Thus, increased education or age will lengthen the waiting time, given the area of search

We now proceed with the formal analysis Let $t(R, E, A)$ be the expected waiting time before a job is found that meets the reservation wage By our preceding discussion

$$\frac{\partial t}{\partial R} < 0, \frac{\partial t}{\partial E} > 0, \frac{\partial t}{\partial A} > 0 \tag{12}$$

We choose the following specific formulation for t

$$t = \frac{\beta(E, A)}{R}, \frac{\partial \beta}{\partial E} > 0, \frac{\partial \beta}{\partial A} > 0 \tag{13}$$

In effect, (13) claims that doubling the area of search will cut the waiting time in half (which is plausible, since doubling the area of search doubles the number of feasible jobs)

Let G be the net benefit per unit of time of switching to another job (which meets the individual reservation wage) The net benefit is measured in utiles and may reflect (aside from the expected pecuniary gains) any expected nonpecuniary gains in pleasure For simplicity, we assume that G is constant over time Let C be the cost per unit of time of not switching jobs, C is measured in utiles and may reflect (aside from the expected pecuniary losses) the nonpecuniary agony of having to keep the unwanted job. Let $f(R)$ be the cost of a job search associated with areas of size R; $f(R)$ is also measured in utiles. Finally, let ρ be the discount rate and $T(A)$ the number of years left to retirement for an individual aged A. Given C, G, and ρ, the present value of utility gained from switching jobs for any area of search R is:

---

[11] The reservation wage is the minimal wage which, if offered, will be accepted by the individual. There is a vast literature on job search which centers on this strategy. See, e.g., Wipple (1973) and Rothschild (1974).

$$U(R, A, E) = G \int_{t(R,E,A)}^{T(A)} e^{-\rho u}\, du - C \int_{0}^{t(R,E,A)} e^{-\rho u}\, du - f(R) \tag{14}$$

Given his expected C and G, an individual of age A and education level E attempts to maximize U by choosing the optimal value for his single control variable R[12] Substituting (13) in (14) and maximizing U with respect to R yields the following first-order condition:

$$(G + C)\,\frac{\beta(E, A)}{R^2} \exp\left[-\rho\,\frac{\beta(E, A)}{R}\right] = f_R(R) \tag{15}$$

Differentiating the LHS of (15) by E yields

$$(G + C)\,\frac{\beta_E(E, A)}{R^2}\left[1 - \rho\,\frac{\beta(E, A)}{R}\right]\exp\left[-\rho\,\frac{\beta(E, A)}{R}\right]. \tag{16}$$

Differentiating the LHS of (15) by R yields

$$(G + C)\,\frac{\beta(E, A)}{R^3}\left[\rho\,\frac{\beta(E, A)}{R} - 2\right]\exp\left[-\rho\,\frac{\beta(E, A)}{R}\right] \tag{17}$$

Note that, if when solving (15) we get $R > \rho\beta(E, A)$, then (16) is positive and (17) is negative[13] This, together with $f_{RR}(R) \geq 0$ (i e , a nondiminishing marginal cost of search), implies (a) that the second-order condition for maximum is met, and (b) that, holding G, C, and ρ constant, increased education E implies an increased area of search R for the first-order condition equation (15) to be satisfied Obviously, if G and C increase with education, the area searched R increases even further. Finally, note that if we differentiate (15) with respect to age A, we get an expression which differs from (16) only insofar as $\beta_E(A, E)$ is replaced by $\beta_A(E, A)$ From (13). the sign of this derivative is the same as that of (16). We established the following proposition

> Proposition 5 Given the cost per unit of time of not switching and the gain per unit of time of switching, the higher the education or age, the larger will be the area searched by a job-seeking individual

---

[12] For simplicity, we assume C and G to be nonrandom The results we obtain will not be altered when uncertainty is introduced if risk aversion is not changed with the education variable

[13] Note that, for $R < \frac{1}{2}\rho\beta(E, A)$, the first-order condition relates to a minimum U In the general job market, the condition $R > \rho\beta(E, A)$ is likely to be satisfied, otherwise (in view of ρ being a small fraction), R, the area of search, may be very small—which is not the case, as is demonstrated by the sizable gross migration. *Note that as ρ approaches* 0, equation (16) will be positive and (17) will be negative for any $R > 0$.

(1973), which uses U.S. Census decennial data; it was found that, ceteris paribus, age does not affect distance elasticity, whereas education increases it significantly. Courchene (1970) obtained similar results using Canadian data

Now, consider the ratio of net migration to gross migration. By the definitions of these measures, the ratio is restricted to the positive unit interval The ratio is zero for any two locations if the flow of migrants in one direction exactly offsets the flow in the other, the ratio is one if the flow of migrants is a one-way flow Generally, the closer the ratio is to one, the more unidirectional is the flow of migrants These ratios reflect the degree of earnings inequality among regions For any two regions, the larger the gap between the earnings-distribution functions (and, thus, between the corresponding medians and the other measures of location), the more the flow of migrants will tend to be one-way in the direction of the economically superior region If two regions are economically identical, the expected flow of migrants in the two directions will offset each other and the net migration will be zero

By propositions 3 and 5, the higher the required education, the larger will be the area searched by each firm or individual, thus, the areas overlap more, generating a marketplace which becomes geographically wider Since response to disequilibrium spreads faster within a market than between markets, the interregional inequality of earnings (for people with similar characteristics) diminishes with education Since the ratio of net to gross migration diminishes as the interregional earnings inequality diminishes, we have

> Proposition 8 The ratio of net to gross migration diminishes with the level of education

Supporting evidence for this proposition is reported in Shryock (1959) and Folger and Nam (1967) It is related to measures of regional earnings inequality; the relation is tested in Schwartz (1971)

V

Rate of Migration, Age, and Education

The verification of proposition 6 is based on interdivisional migration flows in the United States as reported in the 1960 Census [18] The flows are cross-classified by race, sex, returnees-nonreturnees, age, and education. To eliminate variations due to differences in sex roles, racial discrimination, and the different motivations of nonreturnees and returnees, we consider only white male nonreturnees The observed distribution of rates of migration by age and by education can reflect a systematic component —due to the corresponding distribution of the returns to migration by age and education—which may have regional specific component; thus, we have to control for this systematic component

---

[18] In table 8 in U.S. Bureau of the Census, *U.S Census of the Population 1960*, Subject Reports, Educational Attainment, Final Report PC(2)-5B (Washington Government Printing Office, 1963).

The procedure we use is a multivariate regression analysis for each of seven education categories In each category we regress the rate of migration on age dummies and a returns-to-migration variable.

The returns-to-migration variable for each age-education group for each origin and destination is defined as the simple difference between the expected life-span earnings in the destination and in the origin It is constructed on the basis of cross-section earnings profiles, using the median earnings figures for each age-education group in each region These median earnings were obtained from the 1960 Census [19]

The results of these regressions for the seven education groups are summarized in table 1. By adding the corresponding constant to the coefficients of the age dummies, we are able to present graphically the rate of migration (for the interdivisional set of migrants) as a function of age for a few selected educational groups This is done in figure 1

Note that

1 For any educational level, the rate of migration declines with age.

2 The rate of migration for any age increases with education This result is hardly surprising, since the geographical unit of analysis is a census division—and, hence, a long-distance move is implied

3 While the rate of migration increases with education, it declines more rapidly with age as education increases Whereas the rate of migration varies considerably by educational level for the young, this variation is not observable for the older group This finding is exactly what is implied by proposition 6

Note also that·

4 The coefficient of the returns-to-migration variable increases with education Literally, this means that the higher the education, the more responsive are people to the available opportunities; however, a more appropriate interpretation of this finding—in view of the theory developed here—is that the area searched increases with education, which in turn generates more geographic moves

Distance Moved, Age, and Education

The set of migrants to noncontiguous states is a subset of interstate migrants We construct the ratio of noncontiguous state migrants to inter-

---

[19] Table 134 in U S. Bureau of the Census, *U.S Census of the Population. 1960*, vol. 1, *Characteristics of the Population*, (Washington: Government Printing Office, 1963), parts 1–50.

TABLE 1

Age, Education, and the Rate of Migration*

Years of Schooling	R^2	Constant†	Age Dummies‡ 25–29	30–34	35–39	40–44	45–49	50–54	55–59	Returns to Migration
0–4	080	1.89	3.52	2.97	2.31	1.69	1.05	0.683	0.342	0.448
		(0.703)	(0.994)	(0.994)	(0.994)	(0.994)	(0.994)	(0.994)	(0.994)	(0.087)
5–7.	139	2.19	5.96	3.20	2.62	1.67	0.984	0.318	0.091	0.788
		(0.843)	(1.19)	(1.19)	(1.19)	(1.19)	(1.19)	(1.19)	(1.19)	(0.117)
8	214	2.42	7.10	4.69	3.07	1.93	0.992	0.484	0.150	1.19
		(0.818)	(1.16)	(1.16)	(1.16)	(1.16)	(1.16)	(1.16)	(1.16)	(0.126)
9–11	223	2.88	0.796	5.00	3.42	2.08	0.940	0.432	0.187	1.25
		(0.845)	(1.19)	(1.19)	(1.19)	(1.19)	(1.19)	(1.19)	(1.19)	(0.135)
12	263	3.73	7.77	4.89	3.73	2.44	0.901	0.274	0.106	1.50
		(0.779)	(1.10)	(1.10)	(1.10)	(1.10)	(1.10)	(1.10)	(1.10)	(0.142)
13–15	259	4.50	10.6	7.25	6.48	5.17	2.33	0.916	0.666	1.62
		(1.01)	(1.43)	(1.43)	(1.43)	(1.43)	(1.43)	(1.43)	(1.43)	(0.166)
16	.347	5.25	22.0	13.6	9.39	6.98	4.01	1.84	0.435	1.62
		(1.39)	(1.96)	(1.96)	(1.96)	(1.96)	(1.96)	(1.96)	(1.96)	(0.168)

*Dependent variable = rate of migration, standard errors are in parentheses.

†The constant is the contribution to the rate of migration of the reference group, which is the 60–64 group All figures are to be multiplied by 10^{-3}

‡ All figures are to be multiplied by 10^{-3}

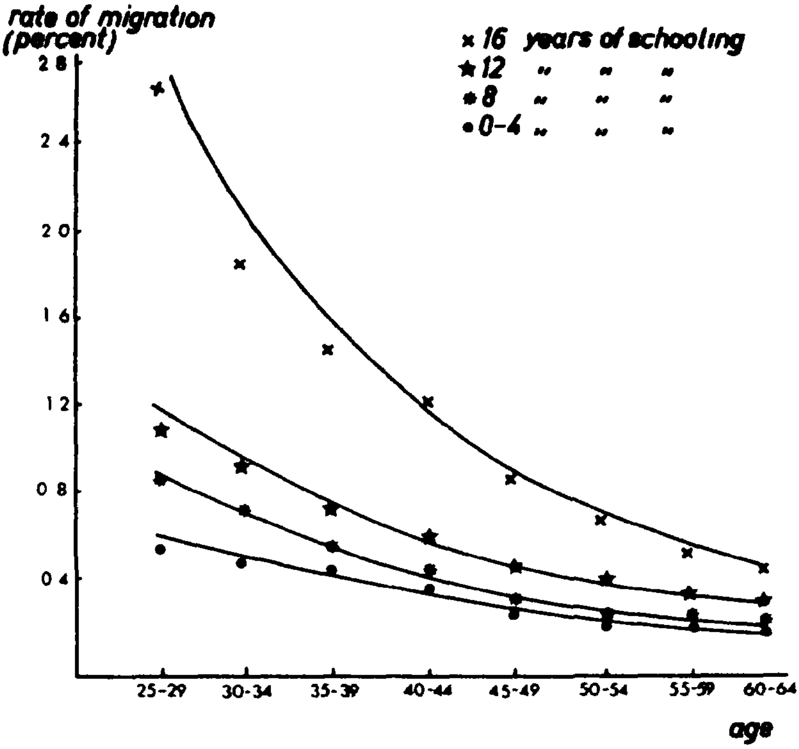


FIG 1—Rate of migration, age, and education

state migrants[20] for each age-education group, starting with the 25–29 age group We regress these ratios on dummies for age groups and education groups The regression results appear in table 2. Note that the coefficients of the dummy variables for education groups increase significantly with education—meaning that the ratio of noncontiguous state to interstate migrants increases significantly with education (which is the implication of proposition 7*a*) In particular, there is a big jump in the ratio for increments of education at the lower educational range From the coefficients of the age dummies, it seems that the ratio declines with age. However, only the ratio of 25–29 year olds differs significantly from the ratios of all other age groups, and by only 5 percent

Another indication of the more important role of education in accounting for the variation of the analyzed ratio is given by the associated *F*-values Whereas the *F*-value of omitting the education classification is

---

[20] The source is table 6 in U S Bureau of the Census, *U.S Census of the Population, 1960*, Subject Reports, Mobility for States and States' Economic Areas, Final Report PC(2)-2B (Washington · Government Printing Office, 1963)

TABLE 2

AGE, EDUCATION, AND THE RATIO OF CONTIGUOUS TO NONCONTIGUOUS STATES' MIGRANTS

	Age			Years of Schooling					
Constant*	25–29	30–34	35–44	5–8	9–11	12	13–15	16	R^2
0.535 (41.49)	0.050 (4.10)	0.010 (0.82)	0.008 (0.68)	0.042 (1.85)	0.110 (7.28)	0.115 (7.72)	0.120 (8.05)	0.130 (8.72)	.91

* The constant is the ratio of the reference group (45–64 years old, with 0–4 years of schooling); *t*-values are in parentheses.

25.1, the *F*-value for omitting the age classification is 6.8. Even though age is not particularly significant beyond the age of 30, the direction of the age coefficients lends support to the claim that transportation cost and its variation with age is not negligible. The increased (with age) transportation cost at least neutralizes the individual's tendency to increase the search area as he gets older.

VI. Conclusion

The shape of the earnings-age function and the way education affects it can explain all the empirically observed relations of migration measures (e.g., rate of migration, net migration) to distance moved, age, and education. No other assumptions—such as attributing lower risk aversion to more-educated persons, or higher psychic cost of moving to older persons—are needed. The individual's search for jobs and the firm's search for employees are affected by the shape of the earnings function—and, thus, by age and education in a predictable way. In turn, the observed systematic behavior of the migration measures with respect to age and education is responsive to the search behavior.

References

Becker, Gary S. *Human Capital: A Theoretical and Empirical Analysis with Special References to Education.* National Bureau of Economic Research no. 80, General Series. New York: Nat. Bur. Econ. Res., 1964.

Brennan, Michael J., ed. "A More General Theory of Resource Migration." *Patterns of Market Behavior*, Providence, R. I.: Brown Univ. Press, 1965.

Courchene, Thomas J. "Inter-provincial Migration and Economic Adjustment." *Canadian J. Econ.* 3, no. 4 (November 1970): 550–76.

David, Paul A. "Fortune, Risk and the Micro-Economics of Migration." In *Nations and Households in Economic Growth*, edited by Paul A. David and Melvin W. Reder. New York: Academic Press, 1974.

Folger, John K., and Nam, Charles B. "Education of the American Population." *1960 Census Monograph, U.S. Bureau of the Census.* Washington: Government Printing Office, 1967.

Greenwood, Michael J. "The Determinants of Labor Migration in Egypt." *J. Regional Sci.* 9, no. 2 (August 1969): 283–90.

Levy, Mildred B., and Wadycki, Walter J. "What Is the Opportunity Cost of Moving? Reconsideration of the Effects of Distance on Migration." *Econ. Development and Cultural Change* 22 (January 1974): 198–214. (*a*)

———. "Education and the Decision to Migrate: An Econometric Analysis of Migration in Venezuela." *Econometrica* 42, no. 2 (March 1974): 377–88. (*b*)

Miller, Edward. "A Note on the Role of Distance in Migration: Costs of Mobility versus Intervening Opportunities." *J. Regional Sci.* 12 (December 1972): 475–78.

Rothschild, Michael. "Searching for the Lowest Price When the Distribution of Prices Is Known." *J.P.E.* 82, no. 4 (July/August 1974): 689–712.

Sahota, Gian S. "An Economic Analysis of Internal Migration in Brazil." *J.P.E.* 76, no. 2 (March/April 1968): 218–45.

Schwartz, Aba. "On Efficiency of Migration." *J. Human Resources* 6, no. 2 (Spring 1971): 193–205.

———. "Interpreting the Effect of Distance on Migration." *J.P.E.* 81, no. 5 (September/October 1973): 1153–69.

Shryock, H. S. "The Efficiency of Internal Migration in the United States." *Proceedings, International Population Conference.* Sponsored by the International Union for the Scientific Study of Population. Vienna, 1959.

Sjaastad, Larry A. "Income and Migration in the United States." Ph.D. dissertation, Univ. Chicago, 1961.

———. "The Costs and Returns of Human Migration." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 80–94.

Stouffer, Samuel A. "Intervening Opportunities: A Theory Relating Mobility and Distance." *American Soc. Rev.* 5 (December 1940): 845–67.

Whipple, David. "A Generalized Theory of Job Search." *J.P.E.* 81, no. 5 (September/October 1973): 1170–88.

Price Dispersion in the Government Securities Market

Kenneth D. Garbade and William L. Silber

New York University

This paper investigates the characteristics of price dispersion in the U.S. Treasury securities market. After a theoretical treatment of why such dispersion exists, empirical evidence is presented which shows that dispersion is influenced by characteristics of securities (such as volume and maturity) and by market supply-demand conditions (as reflected in price-level changes). It is also shown that the cost of liquidity services in a competitive market is determined by price dispersion and is not equal to the bid-ask spread.

The assumption of a unique equilibrium price vector has been enormously useful to economists. This simplification, however, ignores the nature of the exchange process. Stigler (1961) pointed out that for many goods we observe competing inventory specialists who stand ready to buy and/or sell on demand at prices they have posted. These quoted prices will be dispersed over some range, giving an incentive for search by public transactors. Demsetz (1968) observed that the spread between the bid and ask quotes of inventory specialists is the price of the "liquidity" service they provide. These phenomena of price dispersion, search, and bid-ask spreads are alien to a Walrasian world but appear pervasive in our own. Despite the ubiquity of such exchange phenomena, only the subject of

The authors are assistant professor of finance and professor of economics and finance at the Graduate School of Business Administration, New York University. They wish to thank Alan Lerner, Billy Buttner, and their associates in the Government Trading Department at Bankers Trust Company, Tom Smith of Morgan Guaranty Trust, and Al Wojnilower and Dan Donovan at the First Boston Corporation for their detailed comments on the institutions and practices of the market for U.S. Treasury securities. Helpful comments were also received from an anonymous referee. This research was supported in part by the research fund of the Graduate School of Business Administration and the Salomon Brothers Center for the Study of Financial Institutions at New York University.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved.

bid-ask spreads has received significant empirical examination in financial markets (Demsetz 1968; Tanner and Kochin 1971; Tinic 1972; Tinic and West 1974). The empirical characteristics of price dispersion have been investigated by Jung (1960) in the market for automobiles in metropolitan Chicago

This paper is an inquiry into price dispersion in the dealer market for U S Treasury securities. The Treasury market is particularly suited for empirical work, since it includes trading in a wide variety of similar but not identical commodities (issues differ with respect to maturity, amount outstanding, etc). We can, therefore, ask questions about the effects of security characteristics on price dispersion that could not be examined if we were inspecting a market for a single commodity

Section 1 gives a brief description of the Treasury securities market, an outline of the causes and effects of price dispersion and presents the hypotheses to be tested. In Section 2 we describe our data base and the particular part of the Treasury market for which the data are relevant Section 3 presents our empirical results on price dispersion Section 4 presents some estimates of the cost of liquidity services in the Treasury securities market

1. The Dealer Market in Treasury Securities: Structure and Theory

In late 1974 there were 27 primary dealers in Treasury securities (Hawk 1974, p 1) These dealers are inventory specialists who provide markets for the immediate purchase and sale of Treasury bills of almost 40 different maturities and over 50 different Treasury coupon issues Although we are examining here the dispersion of quotations made by these dealers, it is of interest to note that the presence of a dealer market is itself an efficient response to the greater price dispersion which would exist in its absence (Stigler 1961) In addition to the dealers, there are a large number of affiliated participants, including brokerage houses, retail sales houses, and information and communication vendors Meltzer and von der Linde (1960), Hawk (1974), and the studies in the Treasury-Federal Reserve Staff Study (Martin 1969) describe the market and its participants in some detail

In this section we outline a conceptual model of the trading activities of dealers and other investors (defined as all transactors except inventory specialists) that suggests the determinants of price dispersion in the Treasury market We begin with a formal analysis of the behavior of investors who believe that quotation dispersion exists, and who develop search rules as a consequence of that dispersion. We then turn to the behavior of dealers who are cognizant of the search activities of investors and suggest why they quote different prices. We also consider the circumstances under which price dispersion will persist as an equilibrium condition through time. Finally, we summarize why identical transactions can and do occur at different prices at the same point in time, that is, why there is dispersion of transaction prices as well as quotation prices.

a) *Investor Search*

Stigler (1961) suggested that the ignorance of an investor as to the actual condition of offering quotations in a competitive dealer market might be summarized by a probability density function, f_a, such that the probability of seeing an asked price, p (per unit of security), or lower on a given dealer contract is $\int_0^p f_a(s)\,ds$. We assume in this discussion that the investor views all dealers as identical and interchangeable, and thus considers f_a the density function from which the quoted price of any of the potentially contacted dealers is drawn. Let p_0 be the offering price of a particular dealer with whom the investor is in contact. For the size transaction to which our data apply, dealers will usually make their quotes good for some maximum length of time, typically a few minutes, to give the investor an opportunity to check other dealers. If the investor desires to purchase a quantity, V, of securities, his cost to complete the transaction with the dealer quoting p_0 is p_0V. If, however, he samples the offering price of some previously uncontacted dealer, the updated (random) minimum price will be $\min(p_0, s)$, where s is the newly observed quote. Since s is drawn from the density function f_a, his expected cost after the additional contact is

$$C + V\int_0^\infty \min(p_0, s) f_a(s)\,ds = C + V\int_0^{p_0} s f_a(s)\,ds + Vp_0 \int_{p_0}^\infty f_a(s)\,ds, \tag{1}$$

where C is the cost of the additional contact. Stigler (1961) suggested that a major component of C is the investor's time.

The investor will stop searching when his current cost, p_0V, is at least as small as his expected cost after one more contact, implying the existence of a reservation offering price p_a^* which solves the equation

$$C/V = \int_0^{p_a^*} (p_a^* - s) f_a(s)\,ds \tag{2}$$

The investor will accept any quote less than p_a^* and will continue to search if a price at least as low as p_a^* has not been uncovered (Telser 1973).[1]

---

[1] If the investor cannot return to take the offering of a dealer after making additional contacts, eq. (1) does not give the expected cost after one more contact. In the case of such "nondurable" quotes, eq. (1) becomes $C + V\mu_a$, where μ_a is the expected quote from any contact. The corresponding reservation price is $\mu_a + C/V$, which is greater than p_a^* for all quotation distributions.

From (2) it is easy to show that, for given f_a, p_a^* is positively related to the cost of search C and inversely related to the size of the transaction V. It follows that an investor expects to make more contacts the larger the business he has to do and the cheaper the cost of search. The density function f_a need not be an accurate description of the distribution of quotes in the market but only a description of how the investor perceives that distribution. This interpretation differs from that of Stigler but is reasonable in the current context.[2]

For any given perception of quotation dispersion, it is possible to compute the unit price the investor expects to pay to complete his business. This computation permits us to derive the cost of liquidity services in a market with competing dealers quoting dispersed prices. Define

$$PS_a = \int_0^{p_a^*} f_a(s)\, ds \tag{3}$$

as the probability of the investor finding an acceptable offering price on a given contact and

$$\bar{p}_a = \int_0^{p_a^*} s f_a(s)\, ds / PS_a \tag{4}$$

as the expected quotation from a dealer quoting an offer price at least as low as the reservation price p_a^*. Then (as shown in Appendix A) the expected cost per unit is $\bar{p}_a + C/(PS_a V)$ The quantity $\bar{p}_a$ is the price expected to be paid, and $1/PS_a$ is the expected number of contacts required before uncovering an acceptable quote, so that C/PS_a is the total

---

[2] If an investor enters the market only irregularly, his Bayesian prior distribution of prices before making any contacts may be quite flat or diffuse, representing his ignorance of where dealers are quoting For example, suppose the investor knows prices are dispersed according to a stationary distribution parameterized by the average quote μ, $h_a(p, \mu)$, so that

$$\int_0^\infty p h_a(p, \mu)\, dp = \mu.$$

If he knew the average quote, he would know the location of the distribution of prices. Suppose, however, he has only some prior density function g for the average. Then his perception of the distribution of prices would be

$$f_a(p) = \int_0^\infty h_a(p, \mu) g(\mu)\, d\mu,$$

with f_a exhibiting a greater dispersion than h_a if both h_a and g are unimodal radically non-increasing functions. The form of g may be a function of past experience, with its location settled, e.g., by observation of average dealer prices as computed by the Federal Reserve Bank of New York and reported daily in the *Wall Street Journal* for each Treasury issue. After the investor has made several contacts, his information on price quotations improves and we would expect f_a to converge to the true distribution h_a (see Stigler 1961). As f_a is revised upon the acquisition of new information, the reservation price p_a^* will change. Thus, p_a^* is a function of beliefs about conditions in the market as well as objective circumstances.

expected cost of search. A similar analysis of the bid side of the market shows that $\bar{p}_b - C/(PS_b V)$ is the expected per unit revenue from the sale of a volume V of securities net of search costs (where $\bar{p}_b$ is the expected value of an acceptable bid and PS_b is the probability of seeing an acceptable bid on a given contact). The quantities $\bar{p}_b$ and PS_b are defined analogously to (3) and (4). It follows that the expected cost per unit for a round trip transaction, denoted by L, is

$$L = \bar{p}_a - \bar{p}_b + C/V[(1/PS_a) + (1/PS_b)]. \tag{5}$$

In a market where there is only a single inventory specialist, as on the floor of the New York Stock Exchange (examined by Demsetz [1968]), the cost of a round-trip transaction is the bid-ask spread of the specialist plus brokerage commissions, since search, by definition, is not required. For a security traded in a market with competing specialists, however, the expected round-trip cost is the expected transaction spread, $\bar{p}_a - \bar{p}_b$, plus a term which reflects the cost and extent of searching for favorable quotations The expected liquidity cost, L, may not equal the bid-ask spread of any dealer and in general will be a function of the size of the transaction, the perception of price dispersion, and the cost of a contact If the volume of the transaction is sufficiently large, the investor may choose to contact every dealer, so that if there are n market makers the cost of liquidity is

$$\min(p_a) - \max(p_b) + 2nC/V \tag{6}$$

In the limit as V increases, the cost of liquidity services converges to the spread between the *best* quotes on either side of the market. On the other hand, for small-volume transactions or for investors entering the market only infrequently (and for whom the distribution of quotations is diffuse), the expected cost of liquidity may be substantially greater.[3]

b) *Market-Maker Behavior*

We now turn to a discussion of dealer behavior, with the objective of explaining why price quotations on an identical issue may differ. There are five major reasons underlying the disparity between dealer quotations at a point in time. different inventory policies, heterogeneous expectations of future security prices, instability in supply-demand conditions, different cost (of trading) functions, and ignorance of other dealer quotes. These points are all interrelated to some extent, with dealer inventory policies forming the foundation. Each will be discussed briefly in turn.

---

[3] This suggests that those investors who are concerned with the cost of liquidity services, e.g , commercial banks who might have to reverse recent purchases of securities if loan demand proves unexpectedly strong, will, *ceteris paribus*, restrict their investments to issues which trade on narrow spreads and which are characterised by compact dispersions. The larger volume of trading in these issues will tend to further reduce both the spread (Demsetz 1968) and, as we will suggest shortly, the dispersion as well There is simultaneity between trading volume and dispersion as well as between trading volume and spreads.

At any point in time, a market-making dealer will have in mind a desired inventory position The desired level of inventory is a function of its cost of carry (primarily the overnight Federal Funds rate), speculative price anticipations of the dealer, and transactions revenues (bid-ask spread times volume traded). In a market where prices are not changing rapidly, dealers probably prefer a long position (net inventory) to avoid the premium cost of borrowing securities to deliver on short sales. It is also cheaper to sell directly off the shelf than to spend time finding investors willing to lend their securities More generally, active markets encourage larger positions (since delivery must be made on more sales), and higher short-term interest rates discourage larger positions (because cost of carry is higher)

At any moment in time, a particular dealer may find his actual position diverging from his desired position, and he will adjust his quotations accordingly He will typically move his quotes further from what he believes to be the average price quoted by other dealers the less frequent the trading in the issue, since there is a lower probability per unit time of receiving an investor query on such issues Price dispersion stemming from inventory adjustments is likely to be largest, therefore, for inactive issues

Even if desired inventory positions and policies are identical for all dealers, the random arrival of public tenders may cause temporary imbalances which will lead particular dealers to quote aggressively on one side of the market or the other Rothschild (1973) has termed this phenomenon "different market experiences " Indeed, if investors are searching randomly over competing dealers and executing orders at the first acceptable price uncovered, public tenders will arrive randomly at any given dealer Unless a dealer is willing to stand on his quotes, he will not know with certainty whether an adverse inventory experience is a random phenomenon or whether the level of market prices has changed This leads directly to a consideration of expected price movements and instability in supply-demand conditions as a source of price dispersion.

When dealers expect prices to change, they will revise their desired inventory position and alter their bid-ask quotations in order to attain the revised level For example, if a dealer expects bond prices to rise, he will raise his bid and ask in the hope of attracting additional supplies of securities The reverse occurs when bond prices are expected to fall Different expectations among dealers of the direction and magnitude of price movements will lead to quotation dispersion. These differing expectations can stem from alternative interpretations of Federal Reserve behavior or from different information on potential investor activity received from the firm's sales force.

More generally, a change in a dealer's inventory will lead to a change in his quotes both because the dealer wishes to restore his inventory position and because it leads him to revise his estimate of other dealer quotes. Dealers will not, however, maintain markedly different quotes for long, as substantially lower asking prices or higher bid prices will eventually lead to sharp changes in inventory positions, thereby forcing changes in their quotes.

A closely related source of dispersion is instability in market supply and demand conditions (Stigler 1963, p. 181). Dealers continuously search for optimum quotations in light of their expectations When market conditions change significantly, expected future prices change and dealer search for new optimum quotations intensifies Rapid changes in supply-demand conditions lead dealers to quote prices which are characterized by greater dispersion than normal Thus, dispersion may be positively related to the absolute rate of change in security prices This effect will be relatively more important for longer-maturity issues, since the prices of such issues are more volatile

Different cost functions also have an impact on price dispersion. Dealers may quote different prices because they desire to do a greater or lesser volume of business Consider, for example, a dealer whose marginal cost curve (as a function of transaction volume) begins to rise sooner than that of other dealers. This could be because his back-office operations are not as thoroughly automated as those of other dealers or because his capital is smaller, thereby limiting his ability to margin his inventory (Tinic 1972) This dealer can limit his transaction volume by quoting a bid-ask spread wider than that of other dealers, so that he quotes competitively on only one side of the market at any point in time He might change the side on which he is competitive from time to time, according to his inventory experience, but over a given interval he expects to do less business than other dealers

With experience, investors will recognize that such a "one-sided" dealer is, *ceteris paribus*, less likely to quote a favorable price than other dealers and, in order to economize on their own search costs, will accord him lower priority when contacting market makers This could lead to the dealer occasionally losing touch with the flow of public orders. To maintain public interest in his market, the dealer must offer a compensating incentive One such incentive is a reputation for being the highest bidder at those times when he is seeking to buy and the lowest offerer when he is seeking to sell. Such aggressive, albeit uncertain, quotation will lead the dealer to transact inside the prices of other dealers even while he transacts less frequently He can, nonetheless, stay in business if his fixed costs are lower than those of other dealers. Thus, market makers with different cost functions can exist simultaneously. Quotation differences resulting from different spread conventions, and ultimately from different cost functions, will appear to investors as dispersion of quotations by essentially comparable dealers.

Finally, dealers quote different prices because they are ignorant of the quotations of other dealers. In the extreme case, one dealer may be bidding on an issue above another dealer's asked price. Although such opportunities for riskless arbitrage do occur, they also disappear with a speed positively related to the volume of trading in the issue The more trading, and hence search, the greater the probability per unit time that some investor will uncover the arbitrage. Such arbitrage essentially limits the range of dispersion of quotations to the maximum spread quoted by any dealer, since a greater range would imply (for ask prices) that the lowest ask is less than the highest ask less the maximum spread, that is, that it is less than somebody's bid Demsetz (1968) and Tanner and Kochin (1971) found the bid-ask spread inversely related to trading volume. This leads to another reason why quotation dispersion is inversely related to trading volume (since the range of the dispersion is limited by the maximum spread)

c) *Quotation Dispersion, Transactions, and Equilibrium*

We come now to the question of why dispersion persists as an equilibrium phenomenon. Stigler (1961) noted the importance of the rate of obsolescence of information An investor who determines that some dealers are quoting bid prices in a particular order may find this knowledge of little value when he reenters the market at a later point in time Clearly, not every investor could profitably use such earlier experience, since if they could, then the least favorably bidding dealers would find their purchases falling off and their inventory shrinking and they would be forced to bid more aggressively, thus destroying the serial correlation in the bidding order This behavior does not, however, imply convergence by all dealers to a single price, since investor supply and demand schedules are also shifting through time These shifts, when coupled with the different market experiences and expectations of the dealers, lead to continued dispersion of quoted prices

The discussion thus far suggests that price quotations by government securities dealers on identical securities are likely to differ. The existence of search costs implies that otherwise identical transactions can occur at the same point in time at different prices This is precisely what is meant by dispersion of transactions prices. There is no unique market equilibrium price. For very large transactions, it may pay to canvass the entire market, in which case there would be a single market price. But for many transactions and, indeed, for the size of transactions to which our price data refer, the phenomenon of dispersion of transactions prices is likely to correspond closely to dispersion of quotations.

The task that remains is empirical: to identify the factors influencing the distribution of prices in the government securities market. Before we discuss the data, it is useful to summarize the hypotheses about the likely characteristics of price dispersion implied by our theoretical discussion. The dispersion of quotations on each issue should be larger the greater the instability in supply-demand conditions. In particular, price dispersion on days when the level of security prices is relatively constant should be smaller than on days when the level of prices is changing significantly. Similarly, dispersion on any given day should be greater for those securities whose price level is changing most rapidly The characteristics of the individual security should also affect price dispersion. Most prominently, the greater the volume traded in a particular issue, the smaller should be price dispersion We have also suggested that price dispersion (on the bid and ask sides) should be related to the bid-ask spread of each issue In the empirical section that follows, these hypotheses, and additional ones reflecting specific institutional features of the government securities market, are tested

2. Data Sources

Published statistics on financial markets are relatively plentiful and highly accurate compared with (say) data on labor markets or aggregate economic activity In our particular case, however, actual transactions prices on government securities at a particular point in time are not available Individual dealers do not publish their transactions (nor do they publish volume traded) A government securities dealer can call individual dealers and receive "meaningful" quotations, but this is not feasible for a nondealer merely attempting to continuously sample the market without ever consummating a trade

Accurate data on price dispersion are essential for our purposes. Of the 27 dealers in government securities, eight publish a daily quote sheet which lists for each issue a bid and ask price.[4] Such quote sheet prices have been used in studies of the Treasury securities market by Roll (1970) and Fama (1975) If we were to take the written description of the quote sheets at face value, we would conclude that they "give closing markets for marketable Treasury and agency issues"[5] There have been suggestions, however, that the accuracy of quote sheet data leaves something to be desired. It is said that the quotations may not reflect market prices for inactive issues and that the bid-ask spread on the sheets is usually wider than what is quoted to an institutional customer over the telephone

---

[4] Dealers issuing quote sheets are Discount Corporation of New York, First Boston Corporation, Salomon Brothers, Bankers Trust Company, Merrill Lynch Government Securities, Inc., First National City Bank, Morgan Guaranty Trust Company, and Chemical Bank.

[5] See *Handbook of Securities of the United States Government and Federal Agencies* (New York: First Boston Corporation, 1974), p. 135.

Interviews with Treasury security dealers as well as with those responsible for compiling the sheet quotations clearly indicate that the bid-ask quotes are *real* prices at which transactions take place for modest-size public orders, although not for large institutional trades[6] The practice among most dealer firms is to allow their salesmen to pull quotes "off the board" (from which the quote sheet prices are also taken) so as not to tie up traders with transactions smaller than $100,000 for coupon issues and $500,000 for Treasury bills.

Using the quote sheets allows us to evaluate price dispersion in the "odd-lot" market for Treasury securities Among the investors in this market are some private individuals, relatively small banks, thrift institutions, and modest pension and trust funds. The bulk of these investors do not have sophisticated communication links with the dealers, and they rely on ordinary telephone service with low fixed but high marginal costs to search the market They are typically infrequent participants in the market, exhibiting a greater degree of ignorance about recent market conditions than do larger investors They quite likely have more diffuse density functions on price dispersion, and hence there is a strong probability that dispersion of quotations on the sheets of Treasury dealers corresponds to dispersion of transactions prices for this group of investors This was confirmed in interviews with government securities traders and salesmen

Bid and ask prices on 48 Treasury coupon issues (not Treasury bills) from eight government securities dealers were collected for the 22 trading days in January 1975 These data, together with supplementary information, such as the size of the particular issue, maturity, and so on, form the basis for empirically testing the propositions set forth above.[7] The next section reports on the specification and results of those tests.

3. Empirical Characteristics of Price Dispersion

The empirical tests of the hypotheses formed in the previous sections are divided into two parts. First, we consider average measures of the dispersion of bid and ask prices for each issue during the sample month, ignoring intertemporal phenomena. The second set of tests considers daily observations on dispersion and their relation to changes in supply and demand conditions.

---

[6] Interviews were conducted with members of the government securities departments at Bankers Trust, First Boston Corporation, and Morgan Guaranty Trust Company Although trading volume is not available on an issue-by-issue basis, it is available in aggregate by maturity class in the *Federal Reserve Bulletin*. In February 1975, dealer transactions in securities of less than 1 year to maturity were $3.4 billion, between 1 and 5 years, $1.5 billion, between 5 and 10 years, $709 million, and over 10 years, $187 million.

[7] Copies of the quote sheet data are available from the authors. The only item not taken from the sheets is the volume outstanding in each issue, which appears in the *Treasury Bulletin* (U.S Treasury 1975, p. 29).

We investigate the characteristics of price dispersion for Treasury coupon securities with a reduced-form model suggested by the structural interrelationships among trading volume, bid-ask spreads, and price dispersion. The model is formally developed in Appendix B. The model views trading volume, price dispersion, and bid-ask spreads on securities as jointly determined by the exogenous characteristics of the securities in a framework of supply and demand for liquidity services. The exogenous characteristics include: the volume outstanding of the issue (vol); the maturity of the issue (mat), the seasoning of the issue (season), which measures whether the issue is trading actively because it has not yet been "locked away" in investor portfolios, and whether the issue is a "flower" bond (flower), and therefore eligible for payment of estate taxes at par value. Flower bonds are known to trade relatively infrequently.

The reduced-form model is

$$D = \beta_0 + \beta_1 \text{ vol} + \beta_2 \text{ mat} + \beta_3 \text{ season} + \beta_4 \text{ flower} + \beta_5 |\Delta p|, \qquad (7)$$

where D is a measure of dispersion and $|\Delta p|$ is the absolute change in the price of the security. The expected signs are $\beta_1 < 0$, $\beta_2 > 0$, $\beta_3 < 0$, $\beta_4 > 0$, and $\beta_5 > 0$. The formal derivation of these signs is given in Appendix B, but we observe here that the justification is based on the assumption that greater volume of trading in an issue leads to smaller bid-ask spreads and more compact distributions of quotes, as hypothesized in Section 1 above.

a) *Average Characteristics over the Month*

Our initial results on the determinants of the characteristics of dispersion were calculated by averaging daily measures over the sample month. We assume that each bid or ask quotation on a given issue is drawn from a density function whose only source of interday variation is the mean of the distribution. Let $p_a(d, i, t)$ be the offer or ask quote of the dth dealer for the ith issue on day t. We assume $p_a(d, i, t)$ is distributed as

$$p_a(d, i, t) \sim f_a(p, \mu_t, i) \qquad \text{for all } d, \qquad (8)$$

where μ_t is a mean offering quote on day t. In this case the location, but not the shape, of f_a may vary from day to day. The shape may vary with the issue.

The sample statistics we considered were computed for each of the 48 issues and include the average bid-ask spread (S); the n^{th} root of the n^{th} moment about the average quote for $n = 2, 3$, and 4 (Mn_a and Mn_b); and the width of the 90 percent range of quotes (R_a and R_b) about the

TABLE 1

DETERMINANTS OF MONTHLY AVERAGE CHARACTERISTICS OF DISPERSION

Dependent Variable	Constant	ln (mat)	ln (vol)	Season	Flower	R^2	SE
ln ($M2_b$). . .	−2.625 (27.6)	.3715 (7.0)	−.2823 (4.4)	−.377 (2.2)	.313 (2.8)	.850	.269
ln ($M2_a$). . .	−2.841 (30.4)	.5271 (10.0)	−.2208 (3.5)	.643 (3.8)	.273 (2.4)	.893	.264
ln (R_b) . .	−1.371 (14.1)	.3630 (6.5)	−.2672 (4.1)	−.312 (1.8)	.283 (2.4)	.831	.275
ln (R_a) . .	−1.621 (17.9)	.5173 (10.2)	−.2197 (3.6)	−.649 (4.0)	.230 (2.1)	.893	.256
ln (S) . . .	−1.894 (18.6)	.5129 (9.0)	−.1957 (2.9)	−.786 (4.3)	.353 (2.9)	.878	.288

average quote. The subscripts *b* and *a* refer to the bid and ask sides of the market, respectively.

Table 1 shows estimates of model (7) using the standard deviation of quotations ($M2_b$, $M2_a$) and the 90 percent range of quotations (R_b, R_a) as alternative measures of dispersion. The dispersion, maturity, and volume variables are expressed in natural logarithms Maturity is measured in years and volume in billions of dollars of par value. "Season" is a pseudo-dummy variable defined to have value zero if the issue was auctioned earlier than April 1974, unity if it was auctioned in January 1975, and an interpolated value for issues auctioned between those months (larger values for "season" mean the security is "unseasoned"). "Flower" is a dummy variable with unit value if the issue is a flower bond and zero otherwise All coefficients exhibit the anticipated sign and, with one exception, are significantly different from zero at the 5 percent confidence level. The average absolute change in price is not included in table 1. During earlier trials we found this variable highly correlated with the volume and maturity variables. The consequent multicollinearity led to unreasonable estimation results. This was not the case with the daily results, as we will see below.

The results of table 1 confirm our hypotheses on the determinants of dispersion. In terms of relative magnitudes, the elasticity of dispersion with respect to maturity is greater than the elasticity with respect to volume, but the difference is not significantly different from zero for bid quotations. It is significant for ask quotations *Ceteris paribus*, the dispersion of flower bonds is about 35 percent greater than that of issues without the estate tax feature. A new issue has an ask dispersion about 50 percent of what it would be were the issue fully seasoned and a bid dispersion about 70 percent of its seasoned value.

Table 1 also shows the results of using equation (7) to explain the average dealer bid-ask spread on which issues trade (see Appendix B).

The estimated elasticities of spread with respect to maturity (.51) and volume outstanding (−.20) are quite close to the values reported by Tanner and Kochin (1971), .52 and −.28, respectively, in their study of the Canadian government securities market. We estimate that new issues trade on an average spread about 45 percent as large as the spread of similar but seasoned issues and that flower bonds have a spread 40 percent greater than ordinary coupon issues. Comparable results were obtained for a linear form of model (7) and are available on request from the authors. The log-linear version, however, was more powerful in explaining the levels of the dependent variables.

Besides the standard deviation, we also considered higher moments characterizing the dispersion of quotations. The third moment, measuring the skewness of the distribution, is of special interest as a measure of outlying price quotations. Stigler (1963, p 173) hypothesized that ask quotations for reproducible goods are likely to be skewed to the right, since sellers have some minimum but no maximum limit on the prices they can accept. Analogous reasoning suggests that, if Treasury dealers are confident of being able to buy additional inventory at a constant price, offer quotations will be skewed to the right, and if they are confident of being able to sell securities at a constant price, bid quotations will be skewed to the left

We computed the third and fourth moments of the distribution for each issue over the sample month Assuming each distribution is homogeneous of degree one with respect to its standard deviation but otherwise invariant with respect to the exogenous characteristics of the issue, the ratio $M3/M2$ is a standardized measure of skewness and the ratio $M4/M2$ is a measure of whether the distribution is peaked and has fat tails. We found

$$M3_b/M2_b = \underset{(6\ 34)}{-0\ 6225^*} \qquad \text{SE} = 681, \tag{9a}$$

$$M3_a/M2_a = \underset{(1\ 71)}{0\ 1887^*} \qquad \text{SE} = .766, \tag{9b}$$

$$M4_b/M2_b = \underset{(3.71)}{1\ 4638^{**}} \qquad \text{SE} = 276, \tag{9c}$$

$$M4_a/M2_a = \underset{(3.13)}{1.4443^{**}} \qquad \text{SE} = .284. \tag{9d}$$

(One asterisk denotes that the *t*-statistic is on the null hypothesis the coefficient is zero, and two asterisks denote that the *t*-statistic is on the null hypothesis the coefficient is $1.3161 = (3)^{.25}$.)

Bid quotations are skewed significantly to the left, or to lower bids, but ask quotations are not significantly skewed toward higher offering quotes. The evidence implies that dealers are more confident of their ability to sell securities at a constant price than to buy them at a constant price. This asymmetry corroborates the comments of several traders that they are less willing to go short in an issue to accommodate a buyer than to go long to accommodate a seller. Covering a short position, they reason, means they have to locate an owner of a particular issue and possibly offer a premium price to induce him to sell. The demand of a buyer, however, can often be satisfied with any of several issues, implying greater flexibility on the offer side of the market and hence greater confidence by dealers in their ability to sell a particular issue.

The fourth moments suggest that the dispersion of both bid and ask quotes is more peaked and has significantly fatter tails than would characterize quotes normally distributed about the mean quote. When added to equation (9) coefficients on volume, maturity, and the seasoning and flower bond dummy variables displayed statistically insignificant values This justifies the assumption of homogeneity used in equation (9) and implies that the shape of the distribution of quotations about the average quote is determined by the standard deviation characterizing each issue.

b) *Daily Characteristics*

The effect of intertemporal changes in supply and demand conditions on price dispersion can be inspected by applying equation (7), including the term $\beta_5|\Delta p|$, to *daily* observations on the range and standard deviation of quotations Using daily data permits us to separate out, in the regression, observations on dispersion associated with issues that had large price changes on particular days but which, on average, were not especially volatile We expect that the greater the magnitude of the revaluation of the security by the market, the greater will be dealer uncertainty as to the new location of an equilibrium mean quote and hence the greater the dispersion. Table 2 reports regression results for the daily version of the dispersion model Because of the limited number of daily observations on a given issue, the range measure in table 2 is the difference between the minimum and maximum bid (R_b) or ask (R_a) quotes, rather than the 90 percent range as in table 1

The coefficients on the maturity, volume, and flower variables in table 2 correspond closely with the coefficients reported in table 1. The coefficients on the seasoning dummy are somewhat larger than the average monthly results but are still acceptably close The range and standard deviation measures of dispersion in table 2 are computed from a sample of size 8, while those in table 1 are computed from a sample of size 176, so the former are much more susceptible to estimation errors. We expect, and find, that the standard error of the regression rises and the explanatory power of the model falls when daily data are used.

TABLE 2

DETERMINANTS OF DAILY CHARACTERISTICS OF DISPERSION

Dependent Variable	Constant	ln (mat)	ln (vol)	Season	Flower	$\|\Delta p\|$	R^2	SE
ln ($M2_b$) . .	−2.692 (77.88)	.3520 (17 72)	−.2967 (13.87)	− 511 (8 97)	312 (8 17)	.6357 (5.02)	695	.408
ln ($M2_a$) .	−2.872 (83 52)	5188 (26 25)	−.2271 (10 67)	−.840 (14.83)	290 (7.63)	.2727 (2 17)	.767	406
ln (R_b) . . .	−1 643 (43.43)	.3508 (16 13)	−.3058 (13 06)	− 505 (8 11)	.331 (7 92)	.6061 (4.37)	.661	.446
ln (R_a)	−1.838 (49 91)	5289 (24 99)	− 2375 (10 42)	− 836 (13 78)	284 (6.98)	.2471 (1.83)	.747	.434
ln (S)	−1 854 (79 48)	.5165 (38 53)	− 1957 (13 56)	− 819 (21 31)	.328 (12 72)	−.1990 (2.33)	.871	.275

The first four equations of table 2 clearly indicate that, when the price of a security changes a great deal from the preceding day, the dispersion of bid and ask quotations increases significantly. This confirms the hypothesis set forth in the first section regarding instability in supply-demand conditions as a source of price dispersion. On the other hand, the fifth equation indicates that average dealer bid-ask spreads decrease when supply-demand conditions are changing. This is surprising and contrary to our hypothesized relationship. On reviewing our data, we found that the spreads on each issue were remarkably constant during the month which formed our sample period. We had available 384 series of daily observations (48 issues times 8 dealers) on dealer spreads. Seventy-nine percent of those series showed the same spread on every one of the 22 days. Hence, we do not place much confidence in the effect of price changes on spreads reported in table 2. In essence, the negative (and significant) coefficient is a spurious result.[8]

The empirical results of this section substantially support the hypotheses developed in the first section. Dispersion of quotations appears to be inversely related to both the volume of trading in an issue (as reflected in the maturity, volume, season, and flower variables) and the stability of supply and demand conditions. Higher moments of the distribution of quotations appear to be homogeneous with respect to the standard deviation of the dispersion, with some skewness away from the inside quotes and somewhat fatter tails than would be expected from quotations distributed normally.

---

[8] Evidently our sample period did not include any observations on changes in supply-demand that were of sufficient magnitude to lead dealers to widen their spreads. We believe a longer sample interval, which includes sharp breaks in the market, would show a positive effect between $|\Delta p|$ and dealer spreads.

4. The Cost of Liquidity Services in a Competitive Market

In the second section we observed (eq. [5]) that the expected cost of liquidity services in a competitive dealer market is the spread between the expected transactions prices plus the expected cost of search. Both of these elements depend upon the reservation bid and ask prices of the investor, which in turn depend upon the size of the contemplated transaction, the cost of contacting a dealer, and the pattern of price dispersion. In this section we briefly present some calculations on the expected cost of liquidity services for a representative short-term issue (the 5⅞s of February 1976). We assume the investor has perfect knowledge of the average bid and offer quotation prices for the issues and that the density function from which quotations are drawn is equal to the average density function which existed during January 1975.

The top panel in figure 1 shows that, as the size of a transaction increases, the reservation bid and ask prices move closer together The horizontal axis is scaled by volume per dollar cost of contacting a dealer The reservation price depends only on the C/V ratio and not on the absolute size of a transaction, as demonstrated by equation (2) The convergence of the reservation prices implies narrowing of the expected execution spread, but also implies an increase in the expected number of contacts. Thus, the expected cost of liquidity services, shown by curve *LL* in the lower panel, does not decline as rapidly as the expected execution spread.

The lower panel also shows the average dealer bid-ask spread for the issue, line *BA*. As can be seen by the intersection of *LL* and *BA*, for transactions above some minimum level the expected cost of liquidity services from searching the market is less than the average price of such services charged by dealers. Using the spread of a single dealer as a proxy for the cost of liquidity services overstates the social cost of such services as provided by the market as a whole [9] In other words, our results point up the limitations to using the average bid-ask spread as a measure of liquidity costs in a competitive market.

5. Conclusion

This paper tested, in the context of the Treasury securities market, a number of hypotheses concerning the characteristics of price dispersion originally set forth in Stigler's seminal article on the economics of in-

---

[9] We should point out that the reservation prices shown in fig. 1 are probably biased toward inside prices relative to actual reservation prices of investors. The reason is that the reservation prices were computed from eq. (2) as the threshold prices at which the expected gain from an additional contact is just equal to the cost of the contact. However, the cost of the contact is a known cost while the gain is a random variable. This computation of the reservation price ignores the importance of risk in the investor's decision to contact another dealer. This risk increases as reservation prices move toward the inside

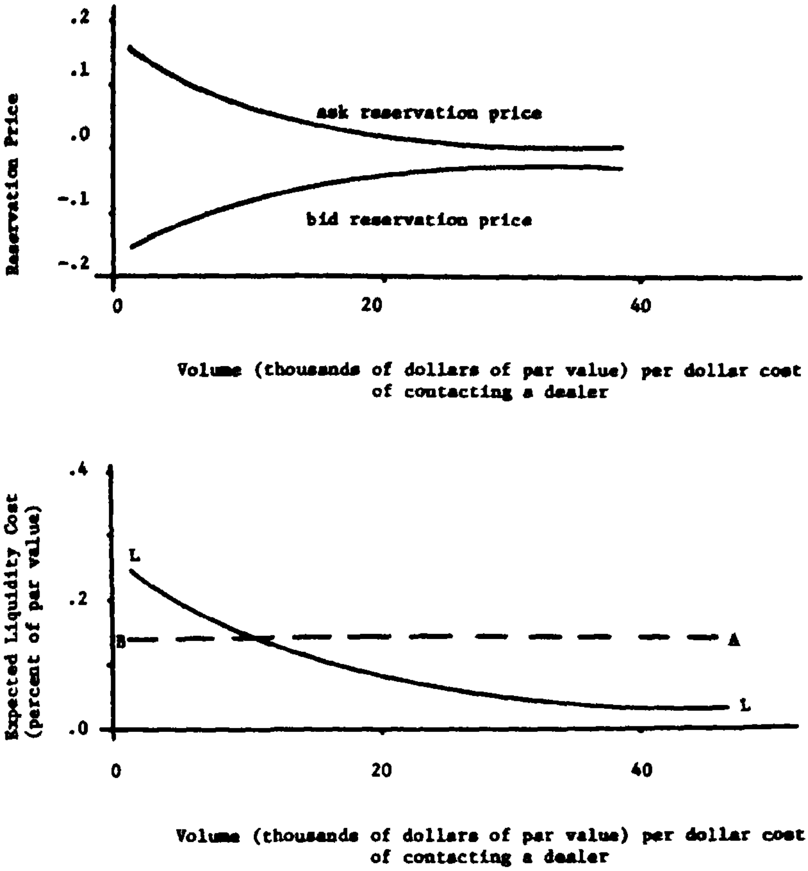


FIG 1 —Relative reservation prices and expected cost of liquidity services for 5⅞s of February 1976

formation The empirical evidence confirms the importance of the theoretical determinants of price dispersion. Even in the sophisticated, well-organized market enjoyed by Treasury securities, the phenomena of search and price dispersion are present and can be explained by existing theoretical models

For those with some lingering skepticism regarding dispersion in such a well-organized market, the following quote indicating price dispersion

---

of the market and the probability of an acceptable contact falls. Depending on the willingness of the investor to assume risk in search, the true reservation prices will be somewhere outside of the certainty-equivalent reservation prices. Thus, fig. 1 overstates the amount of search by investors and understates the expected cost of liquidity services.

in commodities trading—which takes place at a *single* trading post, called a "pit," on the floor of an exchange—may be instructive:

> A third kind of "scalp" is geographical within a pit. Pits get physically congested and when prices are fluctuating sharply . . . a "scalper" may hear a broker on one side of the pit bid a higher price than another broker on the other side of the pit offers. The scalper says "done" to one and "done" to the other. It seems improbable but it can and does happen. [T. A. Hiermonymous, *The Economics of Futures Trading* (New York, 1971)]

Appendix A

The Cost of Liquidity Services in a Competitive Dealer Market

This Appendix derives the expected cost of purchasing securities and the expected costs of liquidity services in a competitive dealer market characterized by dispersed price quotations. The text has already defined the reservation price p_a^* of an investor searching over dealer offerings (eq. [2]), the probability PS_a of seeing an acceptable offer price on a given contact (eq. [3]), and the expected execution price $\bar{p}_a$ (eq. [4]). If we use these definitions, the total expected execution cost is equal to

$$PS_a[C + \bar{p}_a V] + (1 - PS_a)PS_a[2C + \bar{p}_a V] + (1 - PS_a)^2 PS_a[3C + \bar{p}_a V] + \cdots,$$

where $(1 - PS_a)^k$ is the probability of failing to see an acceptable quote in k contacts. Thus, the expected cost is equal to

$$[1 + (1 - PS_a) + (1 - PS_a)^2 + \cdots]PS_a\bar{p}_a V + [1 + 2(1 - PS_a) + 3(1 - PS_a)^2 + \cdots]PS_a C.$$

Now if $V(x) = 1 + x + x^2 + \cdots$ for $|x| < 1$, then $V(x) = 1/(1 - x)$. Since $V'(x) = 1 + 2x + 3x^2 + \cdots$, we have $V'(x) = 1/(1 - x)^2$, so the total expected purchase cost becomes

$$\bar{p}_a V + C/PS_a$$

as stated in the text following equation (4).

Appendix B

Derivation of the Reduced-Form Model for Dispersion

As pointed out in the beginning of Section 3, we investigated the cross-section characteristics of price dispersion with a reduced-form model suggested by the endogenous structural relations among trading volume (T), bid-ask spreads (S), and dispersion (D). We assume that investor demand for dealer liquidity services in a particular issue (reflected in the volume of trading in that issue) is related to the costs of trading, S and D, and the exogenous characteristics of the issue as

$$T = -a_{12}S - a_{13}D + b_1 + c_{11}\,\text{vol} + c_{12}\,\text{mat} + c_{13}\,\text{season} + c_{14}\,\text{flower}, \quad \text{(B1)}$$

where vol is the volume outstanding of the issue, mat is the time to maturity, season is a variable reflecting whether the issue has been auctioned recently (and thus can be said to be unseasoned), and flower is a dummy variable with value 1 if the issue has the privilege of being acceptable for payment of estate taxes at par, and zero otherwise. We anticipate $a_{12} > 0$, $a_{13} > 0$, $c_{11} > 0$, and $c_{12} < 0$ Dealers have commented that issues trade most actively shortly after their auctions, when "season" has its highest value, and decline in trading volume as they are placed in the portfolios of long-term investors, so we expect $c_{13} > 0$. The so-called flower bonds are known to trade less frequently than issues without the estate tax privilege, so c_{14} is expected to be negative

We write the schedule of supply of liquidity services in reverse form,

$$S = -a_{21}T + b_2 + c_{21}\,\text{vol} + c_{22}\,\text{mat} + c_{24}\,\text{flower} + c_{25}|\Delta p|, \qquad \text{(B2)}$$

to emphasize that a given trading volume in an issue leads to a competitive price S for the liquidity services of a single dealer We anticipate $a_{21} > 0$ and $c_{21} < 0$, since dealers can make inventory adjustments more easily on issues which, *ceteris paribus*, either trade more frequently or are outstanding in greater volume (and hence bear less risk in making markets in such issues) We also expect $c_{22} > 0$, since dealers bear more risk, and demand wider spreads, in making markets in longer-term issues (Tanner and Kochin 1971) The quantity $|\Delta p|$ is the absolute value of the change, from the preceding day, in the average dealer quote on the security We expect $c_{25} > 0$, since dealers will demand greater compensation for providing liquidity services in issues where supply and demand conditions are more volatile relative to otherwise comparable issues. Dealers also demand greater compensation in making a market in flower bonds, since the cost of carrying a position in those issues (the current yield less dealer loan rates) is almost always negative, as they bear extremely low coupons. We anticipate $c_{24} > 0$

The reader should note that (B2) is a cross-sectional price schedule which assumes that the aggregate volume of trading in all issues is constant. Should Treasury securities in general become more active for some exogenous reason, we would expect the constant term b_2 to rise in the short run to ration total demand for liquidity services Our sample period is short because our interest is mainly in cross-sectional comparisons, and we are unable to estimate the effects of this industry supply schedule on the spreads of individual issues.

The final structural equation makes dispersion a function of the trading volume, spread, maturity, and the change in the price level,

$$D = b_3 - a_{31}T - a_{32}S + c_{32}\,\text{mat} + c_{35}|\Delta p|, \qquad \text{(B3)}$$

where $a_{31} > 0$ and $a_{32} < 0$ for reasons discussed in the text. Since the prices of longer issues are more volatile than those of shorter-maturity issues, any information known to an investor on relative dealer quotations becomes obsolete faster the greater the maturity of an issue. Stigler (1961) pointed out that persistence of quotation patterns tends to narrow dispersion, since persistence encourages investment in search efforts. We expect $c_{32} > 0$ and, for the same reason, $c_{35} > 0$.

The system of equations (B1), (B2), and (B3) may be written in the structural form

$$\begin{bmatrix} 1 & a_{12} & a_{13} \\ a_{21} & 1 & 0 \\ a_{31} & a_{32} & 1 \end{bmatrix} \begin{bmatrix} T \\ S \\ D \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & 0 \\ c_{21} & c_{22} & 0 & c_{24} & c_{25} \\ 0 & c_{32} & 0 & 0 & c_{35} \end{bmatrix} \cdot \begin{bmatrix} \text{vol} \\ \text{mat} \\ \text{season} \\ \text{flower} \\ |\Delta p| \end{bmatrix}. \qquad \text{(B4)}$$

[illegible] we have to estimate [illegible] $a_{12}a_{21} - a_{11}a_{22} > 0$ [illegible] model.

$$[illegible] + b_2 \Delta M, \qquad (B5)$$

[illegible] and $b_2 > 0$. For a given dealer [illegible] volume relations requires [illegible] volume and spread relations, given [illegible] the indicated signs in (B5) [illegible] of dispersion, the coefficients will [illegible] is quite short, we can safely assume [illegible] and the stability of the coefficients in (B5). Over a longer period, however, the coefficients are quite likely to change

References

Demsetz, Harold. "The Cost of Transacting." *Q.J.E.* 82 (February 1968): 33–53.

Fama, Eugene F. "Short-Term Interest Rates as Predictors of Inflation." *A.E.R.* 65 (June 1975): 269–82.

Hawk, William A. *The U.S. Government Securities Market.* Chicago: Harris Trust & Savings Bank, 1974.

Jung, Allen. "Price Variations among Automobile Dealers in Metropolitan Chicago." *J. Bus.* 33 (January 1960): 31–42.

Martin, William (chairman). *Joint Treasury-Federal Reserve Study of the U.S. Government Securities Market.* New York: Board of Governors of the Federal Reserve System, 1969.

Meltzer, Allan, and von der Linde, Gert. *A Study of the Dealer Market for Federal Government Securities.* Joint Economic Committee Print, 86th Congress, 2d Session. Washington: Government Printing Office, 1960.

Roll, Richard. *The Behavior of Interest Rates.* New York: Basic, 1970.

Rothschild, Michael. "Models of Market Organization with Imperfect Information: A Survey." *J.P.E.* 81, no. 6 (November/December 1973): 1283–1308.

Stigler, George J. "The Economics of Information." *J.P.E.* 69, no. 3 (June 1961): 213–25. Reprinted in *The Organization of Industry* (Homewood, Ill.: Irwin, 1968).

Tanner, J. Ernest, and Kochin, Levis A. "The Determinants of the Difference between Bid and Ask Prices on Government Bonds." *J. Bus.* 44 (October 1971): 375–79.

Telser, Lester. "Searching for the Lowest Price." *A.E.R.* 63 (May 1973): 40–49.

Tinic, Seha. "The Economics of Liquidity Services." *Q.J.E.* 86 (February 1972): 79–93.

Tinic, Seha, and West, Richard R. "Marketability of Common Stocks in Canada and the U.S.A.: A Comparison of Agent versus Dealer Dominated Markets." *J. Finance* 29 (June 1974): 729–46.

U.S. Treasury. *Treasury Bulletin.* Washington: Government Printing Office, January 1975.

e Quantitative Aspects of the onomics Journal Literature

Richard E. Quandt

Princeton University

The paper investigates various aspects of citation practices on the basis of all citations appearing in full volumes of eight principal journals at 10-year intervals. The growth of the number of citations is in good agreement with the findings for other fields, and the mean age of citations declines over time by about 6 months per decade. Most journal articles have a significant tendency to cite other articles published in the same journal, and the cross-reference patterns are employed to rank the journals in terms of importance. An examination of the list of most frequently cited authors reveals that Nobel prizewinners are characteristically on at least one, but most often more than one, list (a decade apart) and permits tentative predictions as to who future prizewinners will be.

1. Introduction

Recent years have witnessed a fair amount of introspective effort among economists This introspection has principally taken the form of various examinations of the journal literature in economics. Most recently, Stigler and Friedland (1975) have examined questions of who cites whom among the Ph D's of certain leading institutions, and Eagly (1975) has treated journal-citation practices as shedding light on the communications network in economics Somewhat earlier, Lovell (1973) estimated production functions for the production of economic articles and also analyzed citation practices from a sample of four journals in 1965 In this work he also discusses the frequency with which certain authors are cited and the age of cited articles, both of these being subjects of interest for the

I am indebted to numerous undergraduate research assistants for combing the journals for the citation data, to Roman Tymiak, Jorge Vizcaino, and Un Chan Chung for various research assistance, to Lester V Chandler and Albert Rees for catching some errors in the raw data, and to Gregory C. Chow and Fritz Machlup for constructive comments. Support from NSF grant 43747X is gratefully acknowledged.

[*Journal of Political Economy*, 1976, vol. 84, no 4, pt. 1]

© 1976 by The University of Chicago. All rights reserved.

present paper. Hanson and Weisbrod (1972) examined, on the basis of the *Index of Economic Journals* (*IEJ*), which economists have produced the most articles (or pages), either in their lifetimes or in the current biennium. Hawkins, Ritter, and Walter (1973) produced a prestige rating of economics journals on the basis of questionnaire data, and Weber (1972) analyzed the journals' acceptance lags and their reasons for rejecting manuscripts. Holt and Schrank (1968), again relying on the *IEJ*, estimated the average growth rate of the journal literature from 1886 on at about 5 5 percent per year.

There is no doubt that all these contributions to the economics or history or sociology of our discipline shed much light on just how it operates They follow in the tradition of the path-breaking work of de Solla Price (1965), who analyzes the incidence of references and the decay of citations with the increasing age of an article, based on citation indices of disciplines other than economics He estimates the growth rate of the world scientific literature at about 7 percent per year and estimates the frequency distribution of articles by the number of citations to them· he finds that 35 percent of all existing articles are never cited, some 49 percent are cited only once, and only 1 percent of all articles are cited six times or more

There are numerous questions one might address to the body of economic literature in existence Among those that are of particular interest in the present paper are the following. (1) How does the "immediacy factor" of de Solla Price (1965) change over time, that is, to what extent is there a change over time in the well-known phenomenon that recent papers are cited more frequently than older ones? To put it in a different way, how does the distribution of the age of citations change over time? (2) What patterns of citation exist among authors publishing in the various journals? (3) Who are the most frequently cited economists?

I shall attempt to answer some of these questions by examining citations in leading economic journals at 10-year intervals from 1890 to 1970.[1] This procedure has, of course, a number of flaws I have, for example, completely omitted an examination of references occurring in books· a common practice in studies of this kind, but nevertheless a serious omission I also take the occurrence or nonoccurrence of a citation as a fact from which important inferences can be drawn with respect to the age of citations, that is, the "memory" of the profession as well as the "fame" of various authors. Yet it is clear that as time passes certain concepts become so intrinsic to a field that their mention no longer requires citation at all. Moreover, intertemporal comparisons of citation frequency of authors must be undertaken only cautiously; since the absolute number of citations grows over time, we may compare citation frequencies over time only in percentage terms or in terms of who is cited most often. There is clearly a possible source of bias in which journals I have selected for examination and a loss of scope in not having identified citations by title but only by author, date, and source of the cited work. I am thus not capable of retracing de Solla Price's study of how often reference is made to a given piece of work.

---

[1] Since various inferences are drawn from the data collected at 10-year intervals, it is fair to ask how sensitive the results might be to the particular years chosen To collect substantially more data would have been prohibitively expensive Since particular attention is paid to citations appearing in 1970, corresponding information from the 1971 issues of the journals in question was also collected, and I will, wherever indicated, compare the 1970 and 1971 results in order to check on the sampling variability of the underlying data In essence it will be seen, as is customarily believed, that *natura non facit saltum*.

Finally, I must comment briefly on the underlying article of faith upon which this and many other similar studies rest. namely, that information of importance is gained about the development of a field and about the significance of the contributors to it by examining the citations in the field's literature To put it in a nutshell, I quote Garfield (1963): "It is preposterous to conclude blindly that the most quoted author deserves a Nobel prize On this basis, Lysenko and others might have been judged the greatest scientists of the last decade The mere ranking by numbers of citations or the numbers of papers published is no way to arrive at objective criteria of importance" (p 290) Obviously these cautionary remarks are well taken Citation frequency does not reveal who will turn out to have been a charlatan and which paradigm is soon to be replaced by another Yet, since it is difficult to arrive at reasoned historical judgments concerning the contributions of recent decades of the quality that, say, Stigler (1965) can confidently make about the eighteenth and nineteenth centuries, I am forced to take citations at face value and hope that serious error survives but briefly

Section 2 deals in detail with my sources of data and the many uncomfortable decisions that had to be made in handling them Section 3 is devoted to the question of the aging of citations, Section 4 to cross-reference patterns among journals, and Section 5 to identifying the most cited authors

2. Data Sources and Principles of Collecting Data

The journals which form the corpus of the study are the following: *American Economic Review* (*AER*), *Journal of Political Economy* (*JPE*), *Quarterly Journal of Economics* (*QJE*), *Economic Journal* (*EJ*), *Economica* (*E*), *Econometrica* (*EM*), *Southern Economic Journal* (*SEJ*), and *Review of Economics & Statistics* (*RESTAT*). Data were culled from all issues of as many of these journals as were in existence in the years 1890, 1900, 1910, 1920, 1930, 1940, 1950, 1960, and 1970. Citations in all articles appearing in these years, referred to as source articles, were recorded, with the following provisos and exceptions: (1) Articles that could be identified as primarily review articles or as works dealing in a major way with subject areas other than economics (such as political science or sociology) were omitted. The former omission was motivated by the expectation that such works would have unusually numerous references, thus being atypical of standard articles in terms of the number of references as well as their overall citation practices. The latter omission was justified by the belief that inclusion of such works would essentially introduce noise into the data. (2) Articles which are comments on recent articles or replies or rejoinders and the like were excluded on the grounds that the age of references contained in them would clearly be biased in favor of the very recent.[2] (3) A reference appearing in an article was counted only once, irrespective of the number of times it was mentioned Thus a reference followed by 20 footnotes containing "Ibid" would all constitute, for purposes of counting references, a single reference to what we shall refer to as a "target journal" (or book or author) References were classified under the eight journals that represent the source journals of the study and under the additional categories of "other journal," "book," and "Ph D. thesis." (4) References to data sources such as various government publications were excluded, as were references to other works written anonymously or issued by governmental or other agencies, commissions, or committees without explicit indication of authorship (5) References were recorded only if they appeared in footnotes and/or in the source article's bibliography References to an author and his work appearing only in text were omitted, on the grounds that to record all such instances would have required an effort of an order of magnitude greater than that required for scanning footnotes. (6) References to authors, even if occurring in footnotes, were omitted if they did not associate the author with a specific piece of work. To illustrate the nature of such references, I cite footnote 2 on page 323 of Clower's article on "Keynes and the Classics: A Dynamical Perspective" (*QJE*, vol. 74 [May 1960]). "It is only necessary to mention such names as Harrod, Domar, Hicks, Modigliani, Samuelson, Goodwin, Duesenberry and Patinkin in order to appreciate the significance of this comment." Clearly, these authors are being referred to primarily because of their written work; but, since there is no indication which of their works are the relevant ones, the references are not useful for measuring longevity of articles (although they could have been useful for measuring how often economists are cited). (7) In some of the early years, references are often given without indication of the year of publication. In these instances reasonable efforts were made to ascertain the proper dates; if this effort failed, the reference was deleted.

---

[2] The data were gathered by undergraduate student assistants. Although they were given reasonably precise guidelines as to how to resolve dubious cases, spot-checking by the author revealed that there are indeed some instances in which I would have decided differently.

3. Growth and Longevity

The basic data are described in table 1 It gives the total number of source articles from which citations were taken and the total number of citations to (1) articles appearing in any of the group of eight journals under examination, (2) all articles published anywhere, and (3) all works including books The journals selected for examination are clearly among those that in some sense may be thought to be important *today* Not all of them existed in all the years reported The compound annual rate of growth of articles in the group of eight is 4 1 percent over the 80-year period This figure is slightly higher than, but not too different from, the growth rate of 3 7 percent implied by Stigler's (1965) figures (p 41) dealing with the total number of articles published in English journals from 1802 to 1853. It is slightly lower than the 5 5 percent growth rate reported by Holt and Schrank (1968) for the period 1886–1963 on the basis of the *IEJ* This latter growth rate is clearly based on a period more comparable with the one used here and also on a more complete enumeration of articles Yet if, as seems plausible, the later decades are characterized by a larger percentage of review articles and "communications" (such as comments, notes, rejoinders, and the like), all of which I have omitted, than the earlier ones, my estimated growth rate would have been closer to theirs were it not for the omission We find that the number of citations of journal articles and the number of all citations grow somewhat faster, at an annual rate of 5 4 and 5.1 percent, respectively This is in complete agreement with de Solla Price's findings, based on Garfield's 1961 Index 2, that the number of citations roughly doubles every 13.5 years (implying an annual growth rate of 5 26 percent) In 1970 my average number of references per article also agrees excellently with de Solla Price's (1965) estimate of 15 for articles published about a decade earlier in various other fields, although Kaplan (1975) notes that this figure is subject to fairly substantial variation among fields On the whole, the number of references per published article has tended to grow [3]

In discussing the "immediacy" factor—that is, the tendency to cite more recently published works more frequently than older ones—de Solla Price observes that 30 percent of all papers cited are between 1 and

---

[3] The comparison of the gross figures for 1970 with those for the year 1971, selected as a check, reveals broad agreement For the latter year the total number of source articles was 438, the number of citations of the eight journals covered here 1,764, and the total number of citations 5,863, giving a mean number of citations of 13.39.

TABLE 1

NUMBER OF SOURCE ARTICLES AND CITATIONS

Year	Number of Source Articles (1)	Citations of the 8 Journals (2)	All Article Citations (3)	All Citations (4)	(4) + (1)
1890	17	14	67	116	6.82
1900	32	31	93	271	8.47
1910	31	23	128	181	5.84
1920	51	67	155	252	4.94
1930	104	141	478	1,043	10.03
1940	163	262	809	1,429	8.77
1950	225	597	1,347	2,076	9.23
1960	291	804	2,250	3,218	11.06
1970	426	1,688	4,414	6,265	14.71

6 years old, the remaining 70 percent being a random sample of all papers ever published. I examine the immediacy factor and its changes over time by analyzing the mean age as well as the median age of citations.

It is clear that, in spite of the attractive aspects of dealing with the mean, caution must be exercised in attaching too much significance to it. A single article dealing with the history of thought published in any one of the years sampled and referring to Aristotle or even St. Thomas Aquinas would inflate the mean unreasonably and would constitute, by all reasonable standards, an outlier. I therefore deal exclusively with trimmed means—that is, means of ages of citations less than K years, where K is a predetermined number. In fact I employ the two values $K = 75$ and $K = 150$. The mean ages of citations and the standard deviations trimmed by 150 and 75, respectively, taking all journals together, are displayed in table 2.

The mean as well as the standard deviations for the individual journals appear superficially to change over time in an essentially haphazard manner. For each of the four series, I regressed the dependent variables (being either a mean age or a standard deviation of age of citations) on time, t, with 1890 equal to 1 and incremented by 1 for each decade. The results are shown in table 3.

The only regression in which the slope comes near to being significant at the .05 level is the one for mean age trimmed at 75 years. According to that regression, the mean age of citations diminishes by about 3 months for every 10-year period. This statistically barely significant finding is, however, inconsequential, since it implies that the diminution of citation age over a century may amount to some 2.5 years.[4]

---

[4] If we consider that subset of citations which consists of references to our basic group of eight journals, we do notice an increase in the mean age over time from means between 7.77 and 9.76 in 1950, 1960, and 1970. This, however, is not surprising, since in the early

TABLE 2

MEANS AND SDs OF CITATION AGES FOR THE AGGREGATE OF JOURNALS

YEAR	MEAN		SD	
	$K = 150$	$K = 75$	$K = 150$	$K = 75$
1890	13.23	12.70	16.53	15.38
1900	10.31	9.04	14.28	10.77
1910	12.01	9.54	17.41	10.03
1920	12.62	10.57	18.98	14.03
1930	13.71	11.51	20.72	14.89
1940	14.99	11.19	22.09	13.45
1950	10.64	9.57	15.13	11.23
1960	11.44	9.77	17.23	11.46
1970	9.67	8.27	15.96	10.72

If we consider the median age of citations, a rather similar picture emerges. From 1890 to 1970, these median figures are 7, 7, 10, 3, 13, 4, 6, 6, and 6 at 10-year intervals. The regression of the median ages on time yields a slope of $-.233$ with an estimated standard error of 0.359, again suggesting an inconsequential decline in age with time that is statistically not significant. We note in conclusion that 50 percent of all references were less than or equal to 6 years old at each of the last three turns of a decade and were less than or equal to 4 years old in 1940—a finding that is again broadly consistent with the more extensive data base employed by de Solla Price.

The fact that there is no particular trend in the mean values over time does not imply that the age distribution for all journals in one year is essentially the same as that in another year. In fact, there are some rather noticeable fluctuations in the median from some one decade to another. The nonstationariness of these age distributions has been tested by performing the Kolmogorov-Smirnov two-sample test on all possible pairs of years. There are 36 such pairs, and, although I well recognize that the Kolmogorov-Smirnov tests for all 36 pairs are not independent of one another, I simply report that the null hypothesis that the age distributions of citations are from the same parent population is rejected in 27 instances.

This suggests that a more disaggregated analysis may reveal some more marked patterns in citation practices. Accordingly, I hypothesized that citation practices, and thus the mean age of citations, may depend not only on time but also on several other variables: specifically, (a) whether the journal is primarily oriented toward econometrics; (b) whether it is British, and (c) the age of the journal. We define the variable y_{it} as the

---

years the group of journals had only a very brief past history—the observed increase in the mean age of citations simply reflects the fact that references to these journals can in the later years have an age distribution similar to the group of all references.

TABLE 3

RESULTS OF REGRESSIONS OF MEAN AGE AND SD OF AGE ON TIME

	Intercept	Slope	R^2
Mean age, $K = 150$	13.005	−.187	.09
	(1.137)	(.202)	
Mean age, $K = 75$	11.478	−.248	.43
	(.719)	(.128)	
SD, $K = 150$	17.163	.086	.02
	(1.741)	(.309)	
SD, $K = 75$	13.668	−.245	.11
	(1.294)	(.230)	

NOTE.—Numbers in parentheses are estimated SEs.

mean age (trimmed at 75) of citations in journal i and year t ($t = 1, \ldots, 9$), x_{1t} as time (t), x_{2i} as 1 for *EM* and *RESTAT* and 0 for all other journals, x_{3i} as 1 for *EJ* and *E* and as 0 for all other journals, and, finally, x_{4it} as the (integer) number of decades of existence that journal i has completed at time t (x_{4it} thus starts with 0 and can become as large as 8). Regressing y_{it} on the x's yields

$$\underset{}{y_{it}} = \underset{(0.98)}{11.91} - \underset{(0.25)}{0.52x_{1t}} + \underset{(1.26)}{0.10x_{2i}} - \underset{(0.88)}{3.47x_{3i}} + \underset{(0.25)}{0.43x_{4it}} \quad R^2 = .37,$$

where the numbers in parentheses are standard errors. Time is now significant, and the decline in the mean age, holding other factors constant, is more substantial, being $\frac{1}{2}$ year per decade Being a British journal is highly significant and takes $3\frac{1}{2}$ years off the mean age. The age of the journal itself is not significant (although nearly so) and has a positive influence on the mean age. This is precisely what one would expect if articles in a given journal had a notable tendency (as I shall establish in the next section) to refer to articles in themselves; for then the younger the journal, the shorter the span of time during which self-references can be made

If we take as our dependent variable the standard deviation of age (again trimmed by 75), we obtain as the regression

$$y_{it} = \underset{(1.26)}{13.88} - \underset{(0.32)}{0.97x_{1t}} - \underset{(1.62)}{0.66x_{2i}} - \underset{(1.13)}{3.92x_{3i}} + \underset{(0.32)}{1.05x_{4it}} \quad R^2 = .46.$$

Only x_{2i} is not significant. The standard deviation diminishes by about a year for each elapsed decade and by almost 4 years if the journal is British, and it increases by a year for each additional decade of life the journal in question has had. This latter conclusion is again consistent with the tendency for journal articles to refer to the same journal, since the longer the past history of a given journal, the greater the dispersion of such "self-references" may be. The fact that a journal's primary orientation is toward econometrics—that is, toward a relatively more mathematical branch of economics—has no significant influence on either the mean or the standard deviation.

The overall conclusion is that (a) growth trends and citation practices in the sample analyzed are in rough agreement with the findings of others for economics as well as for other fields and (b) that, as time passes, the memory span of our discipline, as reflected in the mean or median age of citations, shortens almost imperceptibly and statistically not very significantly, if we take the aggregate of journals, but does shorten more noticeably and significantly if the analysis is disaggregated by journal. This is what one would expect in the light of the increasing volume of scientific literature; the more that is published in the present, the harder it may be to remain aware of the more distant past. Indeed, it would be interesting to find out whether the enormous proliferation of works in recent decades might not have created a greater frequency of ideas being unwittingly rediscovered. Finally, it is also clear that British journals' citation practices are markedly different from those of U S. journals and that the age of the journal itself does have some influence on the age of citation

4. Cross-Reference Practices

We have tabulated the percentage of references in each journal and each year to every journal in the basic group of eight, to books, to other journals, and to Ph.D. theses. The amount of such tabular material is too much to be reproduced here We shall summarize basic tendencies for the early decades in general terms and present the entire cross-reference table for 1970 only (table 4). For several cross-reference patterns not discussed here, the reader is referred to Eagly (1975).

The first general observation is that the fraction of total references accounted for by books is highly variable over time and over journals. For, say, *AER*, this fraction is 58 1 percent in 1890, 76.5 percent in 1900, 25 3 percent in 1920, 54.0 percent in 1940, etc Fairly comparable fluctuations are noticeable for the other journals There is also substantial variation among journals· thus, in 1930 68.9 percent of the references in *JPE* are to books, but only 23.9 percent are to books in *EJ*

The most interesting aspects of cross-reference tabulations are, of course, the implied relations among the eight journals in question. In 1890 there were only a few journals and references, and cross-reference patterns are of no particular interest. By 1900 we have *AER*, *JPE*, *QJE*, and *EJ*. References in *AER* are at this time entirely to books and the category "other journals." Although over 75 percent of references in the other journals in 1900 are also to these two categories, these journals also cite articles in themselves and each other. These references directly to the group of eight journals will be referred to as "group references." References in a journal to an article in the same journal will be referred to as "self-references." Forty-nine percent of *JPE*'s group references are

TABLE 4

Cross-Reference Patterns in 1970

	AER	*JPE*	*QJE*	*EJ*	*E*	*EM*	*SEJ*	*RESTAT*	Books	Other Journals	Ph.D. Theses
AER.	11.3	4 0	3 2	2 0	1 3	3 5	0.3	4 5	26.6	41.6	1.8
JPE. .	5.2	8 4	2 4	3.4	1 0	2 7	0.3	3.3	38.8	32.4	2.0
QJE ..	8.0	7 6	8 1	2 0	2 4	3 4	0 4	2 4	31.3	32.9	0.7
EJ ..	7.2	4 1	4 6	8 5	0 9	1 1	0 0	3.5	39 0	31.0	0.2
E ...	7.5	4.2	1 7	4 5	10 3	6 4	0 0	1 1	26.3	37.7	0.3
EM .	2 2	1 7	1 3	1.2	0 9	15 4	0 1	3.1	26.3	47 1	0.8
SEJ .	9.3	3.3	3 5	2 4	3.3	5 4	2 2	7 1	22.5	38.7	2.2
RESTAT	8.7	6 2	2 3	0 9	0 8	8 1	0 3	10.3	26.8	33.9	1.8

self-references; the corresponding fractions are 72 percent for *QJE* and 71 percent for *EJ*. The picture does not change dramatically in 1910: at this time *AER* begins in a modest way to have group references, and *JPE*'s percentage of self-references rises to 86 percent. Self-references decline proportionately by 1920, although they account in each journal for the single largest percentage of group references. With relatively few exceptions, it continues to be generally true that self-references constitute the single largest percentage of group references. Certain obvious intellectual debts are noticeable, however. In 1930 *E* refers to *EJ* twice as often as to itself, while *EJ* refers to *E* not at all; in 1940 *EJ* refers to itself three times as often as to *E*, while the latter refers to *EJ* somewhat less than that. In 1950 *EJ* refers to itself almost five times as often as to *E*, while *E* refers to itself and *EJ* about equally often. With some variations, this pattern continues to hold for later years as well. The entire cross-reference pattern for 1970 is displayed in table 4, showing the percentage of references occurring in the journals listed on the left to the targets listed on top. It is clear that, within the subset of references we called group references, self-references are the most numerous, with the sole exception of *SEJ*. If citations are taken to be a sign of intellectual debt, then we must conclude that authors tend to submit articles to, and get them accepted for publication by, journals to which their intellectual debt is greatest. Among those seven journals where the self-references are the largest in the set of group references, the fraction of self-references ranges from a low of 23.1 percent for *QJE* to a high of 59.7 percent for *EM*, with all others clustered between 27 and 38 percent.[5]

A final observation that can be made from this table is that one may think of each source journal's reference frequency to other journals as ranking those journals in the group of eight in terms of their importance. If we average the ranks for each target journal (a crude measure, to be sure), we obtain the following mean ranks:

AER	1.14
RESTAT	3.07
EM	3.14
JPE	3.21
QJE	4.00
EJ	4.71
E	5.71
SEJ	7.00

---

[5] Cross-reference patterns are very similar in 1971. There is some variation in the percentages, but no basic patterns change. Self-references still predominate except in the case of *SEJ*. The self-reference frequency for *AER* is 13.6 percent, *JPE* 10.8 percent, *EM* 14.4 percent, etc. The *JPE* is the second most frequently referred to journal by *AER* and conversely, exactly as in 1970. The overall fraction of references in the various journals to journals in the group of eight changes by only a few percentage points. The *SEJ* still refers to *AER* and *EM* most frequently, the relevant frequencies being 8.7 and 5.1 percent. Overall, the agreement between the 2 years is very good.

The reader may form his own conclusions as to whether this rank order is reasonable.

5. Who Are Cited Most Often?

From my data base it was relatively easy to determine who are the writers cited most often. In fact, for each year I determined the 50 most frequently cited authors, irrespective of whether they are cited for journal contributions or books. I nevertheless report only the top approximately 20 of each turn of the decade since 1920. I omit the early years, since the total number of citations is extremely small then, and the number of citations of any one author is negligible, except for the top three or four.

I shall not, however, pass by the early years without at least a brief comment Stigler (1965) presents a table (pp 34–36) of 56 important English economists active in the period 1777–1915 Few would argue with the soundness of his list If we now consider, among the 50 most frequently cited authors in each of the years 1890, 1900, and 1910, those who have been cited more than once in any one of these years, we find that the overlap with Stigler's list is extremely small and extends only to Cairnes, Cannan, Edgeworth, Jevons, Malthus, McCulloch, J S Mill, Ricardo, and Smith (Some non-English economists of note cited more than once include Böhm-Bawerk, J B. Clark, Ely, Fisher, Taussig, Veblen, and von Wieser.)

In table 5 I report the approximately 20 most frequently cited authors for the remaining years The reason for the approximation is that there are numerous ties—I do not want to omit any author with as many citations as another already on the list. The fact that I report approximately the same number of authors for each year implies that the group becomes more and more select as the decades advance, since the number of economists, articles, and citations has grown significantly Being in the top 20 in 1920 is clearly not so great a distinction as being in that group in 1970 For this reason, I biased the list so that it tends to include a slightly larger number of names in the later years. Clearly, it is arbitrary where I cut the lists—nevertheless, a few generalizations may be permitted (1) With the exception of Myrdal and Kantorovich, all Nobel prizewinners appear on the list at least once—most of them two or three times. If frequency of appearance on the lists is any guide toward predicting Nobel prizewinners, the top candidates for a prize in the near future are R F Harrod and J Robinson (3) and then, with two appearances each and in alphabetical order, W. J. Baumol, M. Friedman, G. Haberler, H G. Johnson, L. R. Klein, F. Machlup, R. Solow, and G. J. Stigler.[6]

---

[6] The cutoff points on the list happen to be so selected that, if we included one additional group of names all tied with each other in 1950, one in 1960, and one in 1970, no author who has one mention or none on the present list would advance to the category of

(2) The econometricians compose a very large proportion of the list in 1970, accounting for a third of the names. (3) The only institutions with more than one name on the list for 1970 are the University of Chicago (6), Harvard University (5), the Massachusetts Institute of Technology (3), and Northwestern University (2) These four institutions account for two-thirds of the list in 1970. (4) If a name appears on three or more consecutive lists, a typical pattern (although by no means the only one) is established and is what one would expect. that is, the person starts out low, rises, and then declines Examples are Keynes, whose ranks at four consecutive turns of a decade are 11 (tied), 1 (tied), 2, and 15 (tied), Samuelson, with 2, 1, and 4; and Harrod, 20 (tied), 14 (tied), and 15 This same pattern is in fact evidenced by cases where a name continually rises on the list and then suddenly disappears or starts high, having never before appeared, and sinks steadily (e.g , Hicks, 1, 3, 4, and 25 [tied]). (5) The fields of microtheory, macrotheory, and econometrics account for nearly the whole list for 1970 as well as earlier years. This appears amply to justify the emphasis of graduate curricula on these three areas. (6) Hanson and Weisbrod (1972), in their "economists' hall of fame," list the "current batting champions" in terms of most articles published in each of the four biennia 1960–61, 1962–63, 1964–65, and 1966–67. The union of the four lists includes a total of 51 different names and represents a total of 508 articles (neglecting the possibility of two or more persons on the list having coauthored articles) If our earlier findings are correct, namely that in 1970 the mean age of citations was less than 10 years and the median was 6, then this pool of articles and authors represents the group that has in some sense the highest probability of being cited frequently in 1970 Alas, this is not the case for most of those in Hanson and Weisbrod's list of "current batting champions" The overlap with our list of most frequently cited authors consists only of Balassa, Baumol, Griliches, Johnson, Jorgenson, Modigliani, Samuelson, and Stigler It must certainly be true, as we have always suspected, that the quantity of our output does not ensure that it will be remembered [7]

---

two or more mentions It may be noted that an earlier draft of this paper was written prior to the announcement of the 1975 award of the prize, in that version, Koopmans was among those predicted to receive it

[7] It is relevant to inquire how the rankings and some of the conclusions would have to be modified if 1971 data had been used instead of 1970 data. (1) The 26 most frequently cited economists in 1971 include 21 who made the top 26 in 1970, with J Bhagwati, J M Buchanan, F M Fisher, T C Koopmans, and H. Uzawa replacing five names from the second half of the 1970 list (2) The rank correlation (Spearman's rho) between the ranks of the 21 on both lists is 0 54, significantly different from zero at the 01 level (3) No change is needed in any of the qualitative conclusions above, except that Yale would have joined the other institutions with more than one name on the list and that F M. Fisher would be added to the overlap between the 1971 list of the top 26 and the Hanson-Weisbrod list of "current batting champions" Overall, the agreement between the two sets of results is excellent and sampling variability negligible.

TABLE 5

Approximately 20 Most Frequently Cited Economists

1920		1930		1940	
C. Gini	12	A. Marshall	11	J. R. Hicks*	22
H G. Moulton	9	J M. Keynes	11	J. M. Keynes	17
A. C Pigou	6	H L. Moore	10	F. A. Hayek*	16
W. S. Thomson	6	F. H. Hansen	8	F H. Knight	13
P F. Brissenden	5	A. C. Pigou	8	E Ginsberg	12
A. Marshall	5	L. Robbins	8	A H Hansen	12
W. C. Mitchell	5	F. W. Taussig	8	J. D. Black	11
J. S Nicholson	5	C Davenant	7	R Frisch*	11
E. G. Nourse	5	H. G Moulton	7	G. Haberler	11
F. W. Taussig	5	W. Petty	7	N. Kaldor	11
P H Douglas	4	J A Schumpeter	7	J. Tinbergen*	11
I Fisher	4	J Viner	7	M Ezekiel	10
L. H. Haney	4	N. Allen	6	J M Clark	9
J M. Keynes	4	H. G. Brown	6	P H. Douglas	9
W F Ogburn	4	E Frickey	6	J J Spengler	9
E. R. A Seligman	4	P Wicksteed	6	T. Balogh	8
D A. Tomkins	4	P G. Wright	6	E D. Domar	8
S Webb	4			J Robinson	8
				J Viner	8
				J R Commons	7
				R F Harrod	7
				F. Machlup	7
Total	18		17		22

1950		1960		1970	
J A. Schumpeter	42	P A. Samuelson*	42	K J Arrow*	68
P. A Samuelson*	38	M Friedman	35	M Friedman	68
J R Hicks*	30	L R. Klein	33	R M Solow	53
W C Mitchell	23	J R Hicks*	28	P A Samuelson*	51
A H. Hansen	21	H A. Simon	26	J Tobin	48
L. A. Metzler	21	R. Frisch	25	Z Griliches	41
F. Machlup	17	J Robinson	25	D W Jorgenson	38
O. Lange	16	R M Solow	24	F Modigliani	38
A. Marshall	16	J M Keynes	21	H G Johnson	37
S H Schlichter	15	A Marshall	21	H B Chenery	36
F. H Knight	14	L A. Metzler	20	M Nerlove	33
W W Leontief*	14	W J Baumol	19	L R Klein	31
E D Domar	13	R Dorfman	19	H Theil	29
R F. Harrod	13	G. J Stigler	19	J Johnston	26
E Ginsberg	12	R F. Harrod	18	E S Phelps	26
J M Keynes	12	K J Arrow*	17	A Zellner	26
V Pareto	12	H G. Johnson	17	O Eckstein	25
A C Pigou	12	N Kaldor	17	R Eisner	24
J. Viner	12	S Kuznets*	17	G J Stigler	24
C. Clark	11	A W. Phillips	17	G. S Becker	23
F D Graham	11	F A Hayek*	16	W J Baumol	22
G Haberler	11	T. C Koopmans*	16	J M Lintner	22
T. C Koopmans*	11	J. Tinbergen*	16	B Balassa	21
J Robinson	11			A S Goldberger	21
				G Debreu	20
				J. R Hicks*	20
Total	24		23		26

Note.—Number following each name is citation frequency; asterisks indicate Nobel prize recipients.

6. Conclusion

I have found broad consistency between the growth and citation patterns in the economics-journal literature and that found by others in the same as well as in different fields. I noted that our memories are getting a bit shorter as time passes and that there is a definite propensity for an article to cite the journal in which it is published I determined who are the authors cited most often and found with gratification that Nobel prize-winners are among them but also that the quantity of one's publications does not necessarily guarantee high citation frequency I suspect that there may be a difference in "fame" consisting of having one article cited 100 times and having written 100 articles, each cited only once. It remains, however, for a more ambitious researcher to determine who belongs in the hall of fame by yet another measure—namely, the number of citations per article published.

References

de Solla Price, Derek J "Networks of Scientific Papers" *Science* 149 (July 1965): 510–15

Eagly, R V. "Economics Journals as a Communication Network." *J Econ. Literature* 13, no 3 (September 1975) 878–88

Garfield, Eugene "Citation Indexes in Sociological and Historical Research." *American Documentation* 7 (October 1963) · 289-91.

Hanson, W Lee, and Weisbrod, Burton. "Toward a General Theory of Awards, or, Do Economists Need a Hall of Fame?" *J.P E* 80, no 2 (March/April 1972) · 422-31

Hawkins, Robert G., Ritter, Lawrence S., and Walter, Ingo "What Economists Think of Their Journals" *J P.E.* 81, no 4 (July/August 1973) 1017-32.

Holt, Charles C., and Schrank, William E "Growth of the Professional Literature in Economics and Other Fields, and Some Implications." *American Documentation* 9 (January 1968) · 18–26

Kaplan, Norman "The Norms of Citation Behavior. Prolegomena to the Footnote" *American Documentation* 16 (July 1975) 179–84

Lovell, Michael C. "The Production of Economic Literature: An Interpretation." *J Econ. Literature* 11, no 1 (March 1973) 27-55

Stigler, George J. *Essays in the History of Economics* Chicago. Univ. Chicago Press, 1965.

Stigler, George J , and Friedland, Claire "Citation Practices of Doctorates in Economics" *J P.E* 83, no 3 (June 1975) 477–508.

Weber, John A "Economics Journals Policies, Trends and Problems." *Southern Econ. J* 38, no 4 (April 1972). 559–65

The Origin and Development of Media of Exchange

Robert A. Jones

University of California, Los Angeles

The paper develops an explanation for the emergence of media of exchange through the unconcerted market behavior of individuals. Individuals are assumed to accomplish their ultimate exchanges through trading sequences which minimize the expected time spent searching for complementary trading partners. If individual perceptions of the trading environment are appropriately restricted, then the equilibrium pattern of trade will be some mixture of direct barter and use of a common good as medium of exchange. Although full monetization is always a locally stable exchange pattern, the economy may remain in universal direct barter or partially monetized states.

The general use of media of exchange is one of the most readily observable economic phenomena in exchange economies and represents one of the oldest and most fundamental problems of monetary theory. Media of exchange have been metaphorically described as "oil" in the economic machinery and as "shunting locomotives" in the process of trade (Wicksell 1967, pp. 5, 19), enabling society to escape the intolerable inconvenience, or even impossibility, of finding the "double-coincidence of wants" required for direct barter exchanges (Jevons 1875, p. 3). The mystery and the problem of money were clearly stated by Carl Menger over 80 years ago: ". . . that every economic unit in a nation should be ready to exchange his goods for little metal disks apparently useless as such, or for documents representing the latter, is a procedure so opposed to the ordinary course of things . . . The problem, which science has here to

---

This research was undertaken as part of the author's Ph.D. dissertation at Brown University and was supported by the Canada Council of the Arts. I would like to especially thank Professor Herschel Grossman for his valuable criticism and encouragement. I am also grateful to Professors A. Feldman, M. Parkin, K. Nagatani, and J. Niehans for helpful comments and discussion.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved

solve, consists in giving an explanation of a general, homogeneous course of action pursued by human beings when engaged in traffic, which, taken concretely, makes unquestionably for the common interest, and yet which seems to conflict with the nearest and immediate interests of contracting individuals" (Menger 1892, pp 239–40) The common use of media in exchange suggests a twofold problem · First, what economic gains, if any, accrue from use of a general medium of exchange over and above those accruing from direct trading? Second, how have certain commodities come to be exalted into general media of exchange? This paper addresses the second question It develops a possible explanation for the emergence of media of exchange through the unconcerted market behavior of individuals, without decrees, centralized decisions, or explicit social agreements.

For the purposes of this paper we define a general medium of exchange as the good occupying a particular position in a particular pattern of exchange among individuals Jevons (1875, p 3) described direct barter as "giving what was not wanted directly for that which was wanted," and distinguished it from "sale and purchase in which one of the articles exchanged is intended to be held only for a short time, until it is parted with in a second act of exchange " The good which "temporarily intervenes in the sale and purchase is money " We shall say that an economy has a *monetary pattern of trade* if the exchanges which actually take place have the following two characteristics [1] (*a*) There is one good that enters into every exchange (*b*) Any other good entering an exchange, if purchased is not resold, and if sold is not repurchased The one exceptional good is termed the "medium of exchange " This definition implies that all individuals accomplish their ultimate exchanges through two-stage trades rather than through direct one-stage trades or through longer, more indirect trading sequences The problem is to explain why individuals specifically choose two-stage trading sequences and how all individuals come to choose the same intervening trade good

The literature on monetary exchange has a recurring theme A role for money can be found in the fact that exchange is decentralized, either in the sense that individuals trade with each other rather than with a central "market," or in the sense that there are separate markets for each pair of goods exchanged, and that compensation for goods given up is guaranteed by a *quid pro quo* from the individual or market to which they are delivered. One branch of the literature is concerned with the feasibility of attaining a competitive allocation of goods in a decentralized context Interpreting *quid pro quo* as equal value given up as is received at given market-clearing prices, Starr (1972) points out that Jevon's "double-coincidence of wants" requirement generally precludes fulfillment of excess demands through direct barter. But if there is one good, a medium of exchange, that everyone is willing to temporarily acquire though it is not ultimately desired, then a set of acceptable bilateral exchanges fulfilling all excess demands does exist Ostroy (1973) assumes individuals meet sequentially, cannot give up goods not yet acquired, and cannot base trades on information available only to traders not yet encountered In this context, trading rules which fulfill all excess demands with just one meeting between each pair of individuals are generally incompatible with *quit pro quo* at each exchange Ostroy envisages money as an abstract "record-keeping device," permitting enforcement of budget constraints without requiring *quid pro quo* at every exchange, and thus allowing excess demands to be satisfied with the minimum number of meetings If everyone is endowed with a quantity of one common good sufficient to pay for all their intended acquisitions, Ostroy and Starr (1974) indicate that the less abstract "record-keeping device" of using this good as medium of exchange with *quid pro quo* can achieve the same result Feldman (1973) interprets *quid pro quo* to mean that bilateral exchanges are mutually utility increasing rather than balanced at previously determined prices He proves that the existence of a good which everyone desires and holds in positive quantities, which might be interpreted as money, guarantees that such utility-increasing trades lead the economy arbitrarily close to a Pareto-optimal allocation of goods Although these works illuminate how money might overcome logistical difficulties of reaching an efficient allocation with decentralized exchange, they offer no suggestions for how a monetary pattern of trade could evolve without a centralized decision

---

[1] The monetary pattern of trade may be contrasted with the universal direct barter and trading-post patterns of trade, which also refer to observable regularities in the exchanges that take place in an economy.

The other branch of the monetary-exchange literature focuses on trading costs as determinants of the optimal and equilibrium patterns of exchange Niehans (1969) points out that monetary exchange effects a given reallocation of goods with minimum aggregate transaction costs, subject to *quid pro quo* constraints on trades between individuals, only for particular structures of trading costs Niehans (1971) and Karni (1973) explore the equilibrium trading and consumption decisions made by individuals subject to *quid pro quo* constraints in separate markets for each pair of goods. An appropriate structure of transaction costs can result in a monetary trade equilibrium An objective of this paper is to suggest why the appropriate structure should be so prevalent. Section I describes the general exchange environment in which individuals trade Section II examines how trading decisions might be made Section III investigates the inter-relationships between individuals' trading plans and their implications for the equilibrium pattern of exchange.

I. The Nature of the Exchange Economy

We now make some general assumptions about the economy in which we investigate the equilibrium pattern of exchange. These assumptions concern the trading possibilities and information available to individuals.

1. Voluntary bilateral exchange. Trading opportunities are limited to mutually desired exchanges between pairs of individuals.[2]

2 Exchanges bilaterally balanced at given prices. Individuals are presumed to have some expectations about the terms at which they could make future trades if they acquire a good they do not intend to consume For simplicity we assume that either tradition or a Walrasian "auctioneer" has established consistent prices in terms of an abstract numeraire at which exchanges take place and that no individual will pay more or accept less than these prices for his goods [3]

3 Exchange is costly. We assume that exchange is costly and that each individual regards the costs of different exchanges as fixed and independent of his own trading decisions. More will be said about the nature of these exchange costs below.

4 Limited information about other specific individuals Individuals have no prior knowledge of the endowments, excess demands, or trading plans of other specific individuals. Consequently, they do not know with whom they can or will make given exchanges An individual may have information about other traders as a group, however, based on his past trading experience The four assumptions serve to limit the scope of our problem

The individual's utility function is assumed to depend only on the ultimately acquired bundle of goods, including leisure, and he jointly determines that ultimate bundle and the sequence of exchanges by which it will be acquired. Since the individual takes prices as independent of his own actions, however, the joint solution to the utility maximization problem must have the property that the planned sequence of exchanges acquires the chosen ultimate consumption bundle at minimum total transaction cost. When examining the pattern of exchange we can thus focus on the problem of making a given, fixed ultimate exchange at minimum cost.

---

[2] The Inca civilization of Peru, where the allocation of goods was dictated by the priest class rather than being determined by voluntary bilateral exchange, did not develop any recognizable medium of exchange Neither did many extended family tribes, where the allocation of work and consumption was determined more by consensus and tradition than by bilateral exchange

[3] A *quid pro quo* is required by individuals if they regard their bilateral trade agreements to be independent of one another, and hence the only compensation received for giving up a good must come from its recipient. Bilateral balance at the traditionally given prices could result from individuals believing that future opportunities will always exist for trade at those prices If they ignore the costs of waiting for and finding those opportunities, then their buying and selling reservation prices will equal the traditionally given prices.

The individual does not know a priori with whom he will trade, and in that sense he is "trading with the market." Hence the anticipated costs of a particular exchange of one good for another can depend only on the pair of goods involved and the size of the trade We also note that transaction costs which are additive over the goods exchanged cannot induce the indirect trading when the price of a good is independent of the good for which it is exchanged. Suppose the cost of exchanging a unit value of good i for good j was $c_{ij} = c_i^s + c_j^b$ for all pairs of goods i, j where $\{c_1^s, c_2^s, \ldots\} > 0$ and $\{c_1^b, c_2^b, \quad\} > 0$ are the anticipated costs of handling unit quantities of the goods sold and bought, respectively. Then, since $c_{ik} + c_{kj} = (c_i^s + c_k^b) + c_k^s + c_j^b) > c_i^s + c_j^b = c_{ij}$, direct exchange would always accomplish a simple ultimate exchange at lower cost than indirect exchange. The importance of this point rests with the fact that transportation, storage, and inspection costs could be expected to have this additive form, yet these costs associated with intrinsic physical characteristics of goods are most often emphasized in discussions of media of exchange [4]

What cost of exchange might depend in a nonadditive way on the pair of goods exchanged? An obvious candidate is the cost of finding someone wishing to make the complementary exchange In fact, this is a natural cost on which to focus since classical economists felt that finding a "double-coincidence of wants" was the most inconvenient aspect of direct barter and hence the principle reason why a general medium of exchange is desirable A double-coincidence of desired exchanges is still required for voluntary trade in a monetary economy. An individual must first find someone wishing to trade the medium of exchange for his endowment, then find someone wishing to trade the ultimately desired good for the medium of exchange However, this indirect double-coincidence of desired exchanges may be less costly to find than a direct "double-coincidence of wants"

II. The Individual's Trading Decision

Let us now describe the trading process more closely and examine the pattern of exchange chosen by an individual. Consider an economy with n goods and many individuals, each producing or being endowed with only one good.[5] The good with which each individual is endowed has been previously determined and is fixed. The units in which goods are measured are chosen such that the given prices, in terms of some abstract numeraire, are 1 The only allowable type of trade is the exchange of one unit of one good for one unit of some other good There is no saving, so whenever an individual accumulates one unit of his endowment good he goes to the market and exchanges it for one unit of some good he wishes to consume.[6] The good purchased from trip to trip may vary, however, to achieve any desired average consumption bundle, or to replenish stocks from which the individual consumes at a constant rate.

---

[4] Of course, if the price obtained or paid for a good varies with the good against which it is traded, then a more favorable ultimate price ratio might be obtained through the indirect exchange which could more than compensate for the higher transaction costs. This possibility is made explicit in Niehans (1971). But if the equilibrium price ratios favored the indirect exchange i for k for j over the direct exchange i for j, then they must favor the direct exchange j for i over the indirect exchange j for k for i. If all the additive transaction cost components are strictly positive, then it could not be optimal, in equilibrium, for both the i for j ultimate traders and the j for i ultimate traders to effect their trades indirectly through the same medium of exchange. Consequently fully monetized trade could not be an equilibrium.

To accomplish a given exchange, say of good i for good j, the individual enters the market filled with other traders, but which has no trading posts, and randomly searches for another trader wishing to exchange good j for good i When the complementary trading partner is found the exchange is made, and the individual either continues searching if he has further trades to make or leaves the market if he has accomplished his ultimate exchange.[7] The transaction cost to the individual is the time spent searching for complementary trading partners, and it is this cost that the individual wishes to minimize

As indicated in the previous section, we can assume for our purposes that the individual arrives at the market with a given ultimate exchange to accomplish, even though the trading costs influence which ultimate exchanges are chosen Before entering the market the individual plans a fixed sequence of trades which will effect this ultimate exchange He then enters the market and stays until the planned sequence is executed

Limiting individuals to choosing fixed trading sequences, rather than more general trading strategies which could be conditional upon the types of traders which are first encountered, is a severe restriction From an individual's viewpoint, a conditional strategy can always do at least as well, and usually better, than a fixed strategy The restriction is introduced primarily to allow a tractable aggregation over individual trading plans in Section III

The actual time spent searching for a complementary trading partner is, of course, a random variable We assume that the individual plans his sequence of trades so as to minimize the *expected* time to effect the ultimate exchange Assuming a constant rate of search, he thus chooses a trade sequence minimizing the expected number of encounters with other individuals.[8]

---

[5] We leave unspecified at this point the number of individuals since we will only be concerned with the fractions of individuals having particular characteristics

[6] The individual could come to the market less frequently, say after accumulating five units of his good, as long as he disposed of it through five one-unit exchanges

[7] The restriction to unit exchanges at fixed prices allows all traders to fully complete their currently desired exchange upon meeting the first trader wishing to make the opposite exchange, permitting us to associate the completion of trade with a simple probabilistic event.

The individual bases his trading plans on subjective estimates of the probabilities that a randomly encountered individual would wish to buy or sell any particular good These estimates are derived from his previous experience in the market, and the true probabilities are the fractions of traders in the market waiting to buy or sell the respective goods We now introduce two crucial assumptions about the individual's probabilistic beliefs

1. The individual believes that the same fraction of traders wishes to buy each good as wishes to sell it Since all individuals hold only one unit of a good and demand one unit of some other good, this represents a subjective belief that the market would clear at the given prices It means that the individual's subjective probability that a randomly encountered trader would wish to buy the ith good equals the probability that he would wish to sell the ith good Denote this common probability by p_i

2 The individual believes that what another trader offers in trade and what he desires in return are statistically independent We assume that the individual can estimate the relative commonness of the n goods in the economy, but that excessive computational costs prohibit him from estimating the $n(n-1)$ correlations between goods demanded and offered in trade In keeping with assumption 4 of Section I that he has no prior knowledge of others' trading plans, he assumes no correlation as a working hypothesis Consequently his subjective probability that a randomly encountered individual would wish to exchange good j for good i is $p_i p_j$,[9] and the expected number of encounters he anticipates is required to exchange good i for good j is $1/p_i p_j$ [10]

The above assumptions impose a particular structure on transaction costs as perceived by the individual which greatly simplifies the determination of his chosen trading sequence and which is conducive to monetary exchange [11]

---

[8] Ignoring the variance of time to completion of the ultimate exchange is not critical as long as individuals are either risk neutral or risk averse with respect to the time to completion Since we will characterize the search for a complementary trading partner as a sequence of Bernoulli trials, minimizing the expected time to completion also minimizes the variance of time to completion

[9] The implicit belief that there are traders wishing to trade i for i can be eliminated at the cost of some notational complication, but does not qualitatively affect any of the results

[10] The individual is assumed to be sampling with replacement Another trader could rationally be searched twice since he may have made intervening trades since the last encounter.

[11] The implications of relaxing assumption (2) are substantial The number and variety of equilibrium patterns of trade increase immensely and there is little reason to believe the economy would reach a monetary pattern of trade, at least if the other assumptions are retained.

Consider now the individual's choice of a trade sequence accomplishing the ultimate exchange of one unit of good i for one unit of good j in the least expected number of encounters. If the direct exchange of i for j is chosen, the number of encounters expected is

$$\frac{1}{p_i p_j}. \tag{1}$$

If good j is acquired through a two-stage indirect trade consisting of first trading i for k, then trading k for j, the expected number of encounters required is

$$\frac{1}{p_i p_k} + \frac{1}{p_k p_j}. \tag{2}$$

The individual is free to choose the intermediate good k that he will purchase then resell, and the intermediate good which minimizes (2) is clearly the good with the highest p_k. Let good n be this good. $p_n = \max\{p_k\}$ Thus the minimum expected number of encounters required to effect the ultimate exchange using a two-stage trade sequence is

$$\min_k \left(\frac{1}{p_i p_k} + \frac{1}{p_k p_j}\right) = \frac{1}{p_i p_n} + \frac{1}{p_n p_j} \tag{3}$$

The expected encounters required to complete each stage of the trade are added together since the individual cannot start searching for someone wishing to trade j for n until he has already obtained good n.[12] Also notice that the choice of good n as intermediate trade good is independent of which pairs of goods i, j are being ultimately exchanged.

Similarly, the minimum expected number of encounters that could be anticipated using a three-stage trade sequence, involving two intermediate trade goods, would be

$$\min_{k,l} \left(\frac{1}{p_i p_k} + \frac{1}{p_k p_l} + \frac{1}{p_l p_j}\right) = \frac{1}{p_i p_n} + \frac{1}{p_n p_n} + \frac{1}{p_n p_j} \tag{4}$$

The second trade of the nth good for itself is redundant, of course, and the best two-stage trade sequence will effect the same ultimate exchange in fewer expected encounters Expressions corresponding to longer trade sequences are constructed analogously with the same result There always exists a trade sequence of length 2 dominating all longer trade sequences for each ultimate exchange

Having demonstrated that the individual would never choose an indirect trade sequence of length greater than two, we focus our attention on his decision between direct barter and two-stage trading. Indirect trade could effect the ultimate exchange in fewer expected encounters than the direct trade if and only if

---

[12] Expression (3) would not be appropriate if either negative holdings of goods (credit) were permitted, or if individuals carried excess inventories of commodities for trading purposes, thus allowing them to execute the chosen trades in more than one order. The value of such excess inventories in terms of saved search time can readily be calculated.

$$\frac{1}{p_i p_j} > \frac{1}{p_i p_n} + \frac{1}{p_n p_j}. \tag{5}$$

Multiplying (5) by $p_i p_n p_j$ yields the equivalent condition

$$p_n > p_i + p_j \tag{6}$$

The individual thus chooses an indirect trade sequence of length 2 over direct barter, using the most frequently encountered commodity as intermediate trade good, if he believes that the fraction of individuals in the market desiring to buy or sell that most common good exceeds the sum of the fractions desiring to buy or sell the two goods he is ultimately exchanging.

To the extent that all individuals have the same beliefs about which good has the highest p_k, all indirect traders will use the same intermediate trade good. Since indirect trade in the economy is limited to two-stage trades, all using the same intermediate trade good, we can properly call this nth commodity the medium of exchange However, since (6) could be satisfied for some pairs i, j of goods and not for others, direct barter and monetary trade can consistently coexist. Given our assumptions, any equilibrium pattern of trade will be some mixture of these two extremes

III. The Equilibrium Pattern of Trade

We next investigate the equilibrium mix of direct barter and monetary trade in the exchange economy. Starting from a situation where universal direct barter prevails, and thus where the vector $p = (p_1, \ldots, p_n)$ reflects only ultimate demands for and supplies of goods, condition (6) may be satisfied for relatively few pairs of goods. In other words, indirect exchange may initially be superior to direct barter for a relatively small fraction of the ultimate exchanges. The analysis of this section shows that the medium of exchange will, if used at all, be used much more widely than is apparent from the structure of ultimate demands and condition (6) alone.

This analysis focuses on the determination of the vector p upon which individuals base their trading strategies A component p_i is an individual's estimate of the true fraction of individuals, denoted by q_i, wishing to buy the ith good, either for ultimate consumption *or* as an intermediate trade good. No individual knows why another purchases a good, nor is it relevant to his trading decision. But $q = (q_1, \ldots, q_n)$ will be influenced both by individuals' ultimate exchanges *and* their trading strategies. We define an equilibrium pattern of exchange to be one where beliefs are correct, that is, $p = q$, and where q is consistent with the given ultimate exchanges and optimal trading plans based on p.

The evolutionary process we wish to capture can be described as follows and closely parallels the process envisaged by Menger (1892): Suppose, starting from a state of direct barter, that condition (6) is satisfied for some pairs of goods. The nth good would then be used as a medium of exchange to effect those ultimate exchanges. More individuals would consequently demand, and subsequently offer, the nth good in trade than in the direct barter state. To those demanding the nth good for consumption would be added those demanding it as an intermediate trade good. As this change in the state of the market is gradually revealed to individuals through their sampling of exchange offers during search, they revise upward their subjective estimate p_n (and correspondingly revise downward p_i for $i \neq n$). With the revised p condition (6) could be satisfied for even more pairs of goods, and indirect exchange of these pairs would be chosen over the previously preferred direct exchange. This would further increase the demand for, and supply from subsequent resale of, the nth good and lead to further upward revisions of p_n. An equilibrium is attained when individual beliefs, p, are consistent with the true state of the market. We are particularly interested in how close to full monetization of trade these market forces will take the economy.

Analysis of this evolutionary process requires a dynamic framework for two reasons. First, revisions in p can only take place through continued search and sampling, which takes time. Second, as use of the medium of exchange increases, the fraction of traders offering it in trade will increasingly exceed the fraction demanding it for consumption purposes. The continuous inflow of new traders into the market demanding it as an intermediate trade good makes up this discrepancy between supply and consumption demands. The requirements that the medium of exchange be increasingly common in the market yet remain a relatively small fixed fraction of ultimate demand would be inconsistent with market clearing in a one-period static model. In a dynamic context, however, we can distinguish the stock of traders in the market from the flows of traders into and out of the market and avoid this inconsistency.

To formalize these notions we must specify the distribution of ultimate exchanges, how individual trading decisions are aggregated to determine the state of the market, and a learning process by which individual beliefs embodied in p are revised in accordance with changes in the state of the market. The central variable in which we are interested is the fraction of ultimate exchanges which are executed indirectly through the medium of exchange. We denote this fraction $0 \leq s \leq 1$.

The market is assumed to operate at discrete intervals called "days." Each day a constant number, m, individuals enter the market to make given ultimate exchanges. However, each individual is permitted to make only one trade per day. Those choosing direct barter can effect their ultimate exchange on the day they enter the market, and leave Those choosing indirect trade will acquire the medium of exchange in return for their endowment on the day they enter the market, but return the following day to complete the second stage of their trade sequence. We assume that individuals are indifferent between appearing in the market on just one or on both days and that the relevant cost for their trading decision is still just the total expected number of encounters required when in the market to complete their ultimate exchange. The restriction guarantees that someone will hold the medium of exchange until it is sold to the next group of new entrants

We further assume that all individuals have the same subjective estimate, p, of the state of the market on which to base their trading decisions Since p is based on random sampling, it is in fact a random vector which could differ between individuals [13] For simplicity we use the average p over individuals in each individual's trading decision

A The Distribution of Ultimate Exchanges

Let the matrix $U = [u_{ij}]$ represent the fractions of individuals entering who wish to ultimately exchange good i for good j We assume that U is constant over time and symmetric We define

$$u_i = \sum_{j=1}^{n} u_{ij} = \sum_{j=1}^{n} u_{ji}$$

to be the fraction of new entrants ultimately demanding and supplying the ith good We collectively denote the u_i's by $u = (u_1, \ldots, u_n)$.

The symmetry of U implies that markets will clear at the prevailing prices, in the sense that the same fraction of new entrants demand each good as supply it, and also insures that complementary trading partners can always be found In other words, a double-coincidence of wants is assumed to exist so that direct barter could fulfill all ultimate excess demands if individuals chose that pattern of trade. Furthermore, since condition (6) is symmetric in i and j, and since all individuals have the same p, complementary direct-barter trading partners will always exist even when some ultimate exchanges are conducted indirectly Although this restriction is rather strong, it guarantees that a medium of exchange emerges out of individual choice rather than from necessity.

---

[13] These random differences between individuals' estimated p vectors introduce the possibility that different goods would be used as media of exchange by different individuals. An interesting question is whether the interaction of individual trading plans, random sampling, and expectations revisions would lead one or several media of exchange being used in a stochastic stationary state

The assumption that U remains constant over time is also restrictive. As the state of the market and p change, so does the perceived structure of transaction costs. Hence the relative prices of goods inclusive of transaction costs change, and we should expect some revisions in the excess demands. This model ignores such changes in order to focus more clearly on the changing pattern of trade.

B. Determining the State of the Market

Either direct or indirect exchange will be chosen for a given ultimate exchange depending on whether condition (6) is satisfied or not for that pair of goods by the current p vector. Since u_{ij} is the fraction of entering individuals desiring to ultimately exchange i for j, the fraction s of new entrants choosing indirect trade is

$$s = s(U, p) = \sum_{\substack{i,j \text{ such that} \\ p_n > p_i + p_j}} u_{ij} \, . \tag{7}$$

Since condition (6) cannot be satisfied if either $i = n$ or $j = n$, the variable s always lies between 0 and $1 - u_n$. The situation $s = 0$ corresponds to a state of universal direct barter, and $s = 1 - u_n$ corresponds to complete monetization of exchange.

The true state of the market need not coincide with individual beliefs p about the state of the market. The vector $q = (q_1, \ldots, q_n)$ denotes the true fractions of individuals in the market demanding the respective goods in trade. The total number of traders in the market on a given day is $(m + ms)$, where m is the number of new entrants and ms is the number of traders from the previous day returning to complete their indirect exchange.

The number of individuals demanding the nth good is $(mu_n + ms)$, where mu_n represents the new entrants demanding it for consumption purposes and ms represents the new entrants demanding it as an intermediate trade good. Hence

$$q_n = \frac{mu_n + ms}{m + ms} = \frac{u_n + s}{1 + s} \, . \tag{8}$$

The s in the numerator properly belongs to the current day, while the s in the denominator actually belongs to the previous day. Although s may change substantially over long periods, we assume that it does not change significantly between two consecutive days, so that (8) is a close approximation to q_n at each moment in time.

The number of traders demanding any good i other than the medium of exchange is mu_i. They include the new entrants choosing to acquire the ith good by direct barter and the previous day's entrants who chose to acquire the ith good by indirect trade. To the extent that p, and hence trading strategies, have not changed significantly from the previous day, the number choosing indirect trade yesterday to acquire i will equal the number choosing indirect trade in the current period to acquire i. Thus the total number demanding i on a given day will approximately equal the number of new entrants who ultimately demand i, namely, mu_i. Consequently

$$q_i = \frac{mu_i}{m + ms} = \frac{u_i}{1 + s} \qquad \text{for } i = 1, \ldots, n - 1. \tag{9}$$

The vector q thus constructed also represents the fractions of traders offering the respective goods in trade The number offering the nth good is $(mu_n + ms)$, where mu_n represents new entrants with endowments of n, and ms represents traders who acquired good n the previous day as a medium of exchange For other goods, mu_i is the number of new entrants endowed with good i [14] Equations (8) and (9) can be used to write q as a function of u and s in vector form·

$$q = q(u, s) = u + \left(\frac{s}{1 + s}\right)(e_n - u) \tag{10}$$

The vector $e_n = (0, 0, \quad , 0, 1)$ in the above expression

C The Learning Process

The commonly held beliefs p about the state of the market are assumed to be revised in accordance with the true state of the market by the simple adaptive formula

$$\Delta p^t \equiv p^{t+1} - p^t = \lambda^t(q^t - p^t) \qquad 1 > \lambda^t > 0. \tag{11}$$

Use of this formula is suggested by the optimality of the analogous-expectations adjustment formula for individuals under certain conditions,[15] and by the fact that q^t is the expected value of a random sampling of the state of the market at time t.

---

[14] We have implicitly assumed through our formulation of the q vector that the stock of traders in the market is in equilibrium, or close to it, at all times. As long as the pattern of trading does not change, the flow equilibrium of traders into and out of the market will maintain the stock equilibrium But when individuals first switch from making a given ultimate exchange directly to making it indirectly a stock disequilibrium results. There are not enough units of the medium of exchange commodity in the market to satisfy the demand, and there are too many units of the goods now to be exchanged indirectly. In the process of moving from direct barter to full monetization, $m(1 - u_n)$, additional units of the medium of exchange commodity must be accumulated in the market and consumption of the same number of units of the other goods combined is shifted forward by 1 day For this to occur, there must be some days when $u_{nt} > u_{it}$ for goods $i \neq n$ We are assuming that forces not formally embodied in the model, slight price changes inducing output or consumption changes, cause appropriate temporary disequilibrium flows into the market (changes in U) to correct these stock disequilibria as they occur If monetization takes place only slowly, then, since the total stock adjustment during the whole process is less than even 1 day's trading volume, we can expect the disequilibrium flows to be slight. Expressions (8) and (9) based on a fixed U remain valid as approximations. A more explicit analysis of this process is clearly needed, but is beyond the scope of the current model.

The behavior over time of our dynamic variables s, p, and q is now completely determined by equations (7), (10), and (11) together with the parameters U and an initial value for p. To start the system we assume that the economy has been in a state of universal direct barter for all previous time. Since $s = 0$ implies $q = u$, we choose $p^0 = u$ as the initial value for p.

D. Evolution of the Pattern of Exchange

The dynamic system represented by (7), (10), and (11) is difficult to analyze in its current form since there are $2n + 1$ dynamic variables in p, q, and s, and since the functional dependency of s on p is discontinuous. However for our initial value of $p^0 = u$ we can reduce the system to an equivalent system with only two dynamic variables and proceed with a graphical analysis

In equation (10) notice that the vector-valued variable q depends only on the scalar s and fixed parameter u. As s varies between 0 and $1 - u_n$ the vector q moves along a one-dimensional line segment in R^n with endpoints u and $u + (e_n - u)(1 - u_n)/(2 - u_n)$. Our initial value for p is u, and hence is on the same line segment to which q is confined The adjustment equation (11) for p implies that p is always moving directly toward the current value of q Since p starts on the same line segment as q in R^n, it too will always remain confined to that line. This fact allows us to summarize the $2n$ variables, p and q in two scalar variables, indicating their respective positions along the line segment.

More formally, we introduce a new variable, s^e, which will locate p along the line segment exactly as s locates $q = q(u, s)$. Supposing that

$$p^t = q(u, s_t^e) = u + \left(\frac{s_t^e}{1 + s_t^e}\right)(e_n - u), \tag{12}$$

then

$$\begin{aligned} p^{t+1} &= p^t + \Delta p^t = p^t + \lambda^t[q(u, s_t) - p^t] \\ &= (1 - \lambda^t)q(u, s_t^e) + \lambda^t q(u, s_t) \\ &= u + \left[\frac{(1 - \lambda^t)s_t^e}{1 + s_t^e} + \frac{\lambda^t s_t}{1 + s_t}\right](e_n - u) \qquad (13 \\ &= q(u, s_t^e + \Delta s_t^e) \\ &= q(u, s_{t+1}^e). \end{aligned}$$

---

[15] Specifically, if an individual revises his beliefs about the proportions of individu wishing to sell various goods in a Bayesian fashion, regards his search as sampling from multinomial distribution with unknown parameters $\beta = (\beta_1, \ldots, \beta_n)$, has a Dirich

The new variable s^e may be interpreted as follows: If individuals believed that s^e was the fraction of individuals choosing indirect exchange, and if they knew the value of the parameter u, they could deduce the state of the market to be $q(u, s^e)$. It must be emphasized, however, that individuals in this economy do not actually form expectations about either u or the extent of indirect trading. They only form expectations about q without presuming rationality on the part of any other traders

An adjustment equation for the variable s^e can now be derived. From the definition of $q(u, s)$ in (10) and the third and fourth lines of equation (13) it is clear that, suppressing the time subscripts,

$$u + \left[\frac{(1-\lambda)s^e}{1+s^e} + \frac{\lambda s}{1+s}\right](e_n - u) = u + \left(\frac{s^e + \Delta s^e}{1 + s^e + \Delta s^e}\right)(e_n - u). \tag{14}$$

A straightforward manipulation allows us to solve (14) for Δs^e.

$$\Delta s^e = \left[\frac{\lambda(1+s^e)}{1 + (1-\lambda)s + \lambda s^e}\right](s - s^e). \tag{15}$$

Thus the variable s^e changes according to an adaptive scheme with an adjustment coefficient which varies between 2λ and $\lambda/2$ depending on the current values of s and s^e, and equaling λ whenever $s = s^e$

The final step in the transformation is to specify s as a function of U and s^e alone Equation (7) defines $s(U, p)$ and (10) defines $p = q(u, s^e)$. Substitution of (10) into (7) yields

$$s(U, s^e) = \sum_{\substack{i,j \neq n \text{ such that} \\ u_n + s^e > u_i + u_j}} u_{ij} \tag{16}$$

Evolution of the pattern of exchange can now be investigated through the transformed system

$$\begin{aligned} s_t &= s(U, s_t^e) \\ \Delta s_t^e &= s_{t+1}^e - s_t^e = \alpha_t(s_t - s_t^e) \\ s_0^e &= 0 \end{aligned} \tag{17}$$

We assume that λ in (11) is less than $\frac{1}{2}$ so that α_t always lies between 0 and 1 This insures that s will not jump over one locally stable equilibrium and move on to another. The initial value of 0 for s^e indicates that direct barter was used prior to time 0. Time 0 may be thought of as the time when individuals first considered trading indirectly

---

prior distribution on p with mean p^t, and takes q^t as the sample proportions in a finite sequence of multinomial trials (i e , a market day), then the mean p^{t+1} of his posterior distribution on p is given by (11). The scalar λ^t depends on the precision of his prior distribution and the number of trials in q^t. It is not important for our purposes that λ^t may vary from day to day.

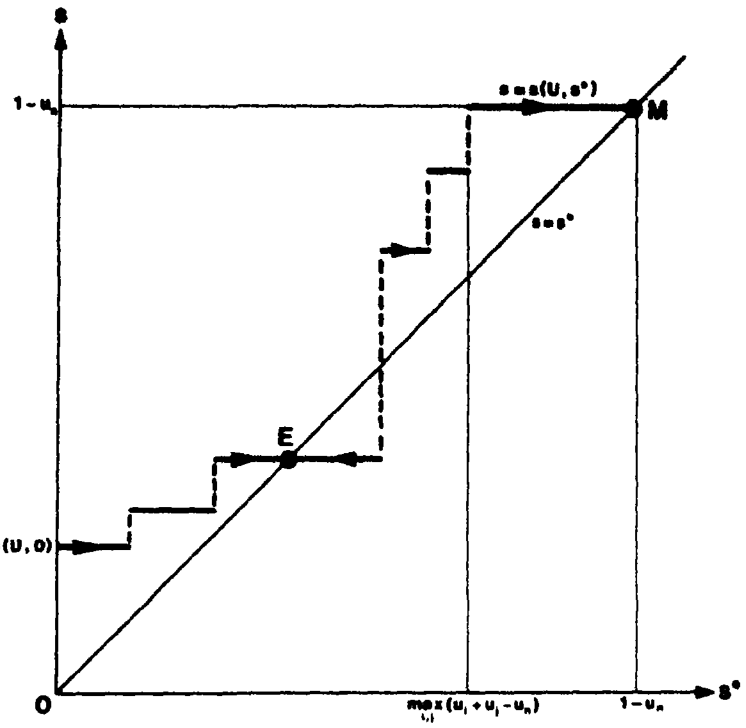


FIG. 1

System (17) is analyzed graphically in figure 1 The variables s and s^e will always be on the graph of $s(U, s^e)$ From its definition in (16), it is clear that $s(U, s^e)$ is an increasing step function in s^e and can only take on values between 0 and $1 - u_n$ It steps up whenever s^e increases sufficiently for indirect trade to be chosen for still another ultimate exchange. Furthermore, $s(U, s^e)$ attains the value $1 - u_n$ whenever s^e exceeds $u_i + u_j - u_n$ for all $i \neq n, j \neq n$.

The adjustment equation for s^e implies that s^e moves α_t of the distance from the 45° line $s = s^e$ toward that line each period. With α_t between 0 and 1, the variable s^e will never cross that line but will always move toward it. Thus a locally stable equilibrium combination of direct barter and monetary trade exists wherever $s(U, s^e)$ crosses $s = s^e$ horizontally. In the example of figure 1 both points E and M are stable equilibria. However if we start with an initial value of $s^e = 0$ the pattern of trade will only reach point E, where direct barter coexists with use of the medium of exchange. By suitably choosing the parameter matrix U of excess demands, examples can be constructed with stable equilibria at the origin, several intermediate points, and at $s = 1 - u_n$. Figure 2 depicts

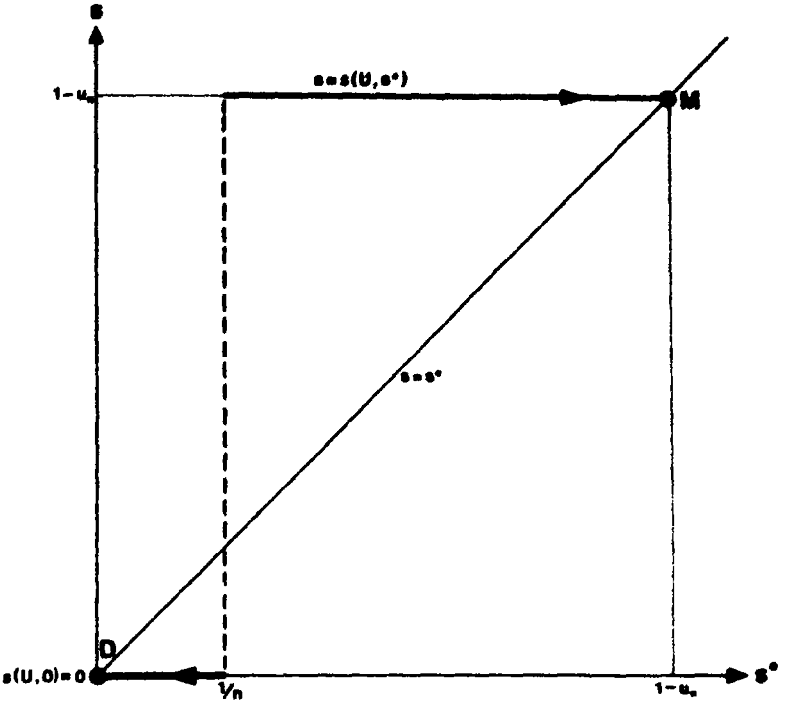


FIG. 2

the case where $u_i = 1/n$ for all i, in other words where all goods appear equally frequently in ultimate exchanges. Only the points D and M, corresponding to universal direct barter and full monetization of trade, are stable equilibria. The economy would stay in the direct barter state in which it begins.

Several dynamic properties of the pattern of exchange are now apparent. First, there may be several stable equilibrium levels of monetization of trade. Even for a single given distribution of ultimate exchanges, direct barter, full monetization, and intermediate mixes may all be locally stable situations. Although we do not develop here any criterion for evaluating alternative patterns of exchange from a welfare standpoint, multiple stable equilibria indicate that decentralized market forces such as we have considered cannot be relied upon to lead the economy to any socially optimal level of monetization. The economy could remain in a state which imposes greater total transaction costs than some other, possibly stable, state. This result is not surprising given the externalities in transaction costs present in the model.

Second, we note that full monetization, $s = 1 - u_n$, is a stable equilibrium for all possible values of U. Thus if an exogenous shock either to expectations p or to ultimate demands U once caused any good to be accepted as a general medium of exchange, then its use would continue even after the shock had passed. Full monetization is the only stable equilibrium pattern of trade, and hence a globally stable equilibrium, if and only if the graph of $s(U, s^e)$ never lies below $s = s^e$, or equivalently, if and only if

$$\sum_{\substack{i,j \neq n \text{ such that} \\ u_n + s^e > u_i + u_j}} u_{ij} \geq s^e \qquad \text{for all } 0 \geq s^e \geq 1 - u_n \tag{18}$$

Third, starting from an initial direct barter state of $s^e = 0$, we require that $s(U, 0) > 0$ if the economy is ever to move away from that state. If $u_n < u_i + u_j$ for all triples of goods n, i, j, then the direct barter state continues as a stable equilibrium Some initial dispersion in the relative "commonness" of goods is required for the monetization of trade to begin

Finally, returning to the initial point to be made in this section, we note that if $s(u, 0) > 0$, in other words, if indirect exchange takes place at all, then the equilibrium level of s generally will exceed $s(U, 0)$ If a medium of exchange is used at all, then it will be used more widely than would be apparent from the vector u of ultimate demands and condition (6) alone

IV. Conclusions

The use of search costs to guide individuals through a market with neither trading posts nor clearing houses has yielded plausible implications for the pattern of exchange With certain assumptions about the costs of finding traders with whom a desired exchange could be made, we have demonstrated why an individual may prefer to execute a given ultimate exchange indirectly rather than directly Two trades can be less costly than one However, given that we have opened up the possibility of lengthy indirect trade sequences, we have also demonstrated why sequences of more than two trades would not rationally be chosen. Three or more trades are always more costly than two. Further assuming that all individuals form their beliefs about market parameters by sampling from the same market, we have an explanation for how all traders could independently, without centralized decisions or agreements, settle on the same good as a medium of exchange Thus individual optimization implies that the pattern of trade must always be some combination of direct barter and monetary exchange.

The particular combination of these two extreme patterns of exchange to which the economy converges depends on the specific configuration of ultimately desired exchanges and on the initial pattern of trade. There can be multiple locally stable equilibrium combinations, but full monetization is always stable.

And finally, the approach suggests that a very common good would emerge as a first commodity money in a barter economy[16] The important point is that this commonness is a market characteristic of goods rather than an intrinsic physical characteristic such as portability, divisibility, or cognizability. This is not to say that such physical characteristics play *no* role in determining *which* good will be used as a medium of exchange. However the analysis suggests that the rationale for using a medium of exchange in the first place might be found in the differing market characteristics of goods and the decentralized nature of exchange

It need hardly be pointed out that these results were obtained in a very specialized context Individuals were compelled to trade at given prices, not permitted to choose from more efficient conditional trading strategies, and formed their expectations about the likelihoods of various trading opportunities in a most simplistic manner Further work relaxing these restrictions appears warranted to determine whether the phenomenon of monetary trade is indeed most naturally a consequence of individuals' limited information about each other's trading intentions

References

Einzig, P *Primitive Money in Its Ethnological, Historical and Economic Aspects* London Eyre & Spottiswoode, 1949

Feldman, A M. "Bilateral Trading Processes, Pairwise Optimality and Pareto Optimality " *Rev Econ Studies* 40 (October 1973) 463–74

Jevons, W S *Money and the Mechanism of Exchange* London Appleton, 1875.

Karni, E "Transactions Costs and the Demand for Money " *Western Econ. J.* 11 (March 1973) 71–80

Menger, K "On the Origins of Money " *Econ J* 2 (June 1892) 239–55.

Niehans, J "Money in a Static Theory of Optimal Payment Arrangements." *J. Money, Credit and Banking* 1 (November 1969) 796–826

——— "Money and Barter in General Equilibrium with Transactions Costs." *A.E R.* 61 (December 1971) 773–83

Ostroy, J M "The Informational Efficiency of Monetary Exchange." *A.E.R.* 63 (September 1973) 597–610

Ostroy, J M , and Starr, R M "Money and the Decentralization of Exchange." *Econometrica* 42 (November 1974) 1093–1114

Starr, R M "The Structure of Exchange in Barter and Monetary Economies." *Q.J E* 86 (May 1972) 290–302

Wicksell, K *Lectures on Political Economy*. Vol. 2 *Money* New York. Kelley, 1967.

---

[16] This suggestion that a very common good would emerge as a first-commodity money does not appear inconsistent with casual anthropological evidence In relatively primitive and decentralized societies either a staple consumption good, principal export item, or traditional obligatory gift frequently emerged as an early medium of exchange Examples include rice in Japan, beaver pelts in Canada, tobacco in the southern American colonies, goats in British Africa, and butter in fourteenth-century Norway. Other examples can be found in Einzig (1949)

Alternative Stabilization Policies for Less-developed Economies

---

Basant K. Kapur

University of Singapore

This paper presents a model of price stabilization applicable to less-developed economies. Distinctive features of the model are (i) its explicit incorporation of a commercial banking system, (ii) its specification that the actual rate of inflation depends on inflationary expectations and on excess demand for output, and (iii) endogeneity of the growth rate of real output. Numerical simulations are also conducted. We demonstrate that stabilization through an initial increase in the average nominal interest rate paid on money holdings has significantly more favorable short-run effects on real output than does stabilization through an initial reduction in the rate of monetary expansion.

The short-run dynamics of price stabilization in inflationary economies has been explored in a number of theoretical studies (see, e.g., Mundell 1965, Friedman 1970, Selowsky 1973, and Frenkel 1975). As has been emphasized by Shaw (1973) and McKinnon (1973) in two recent works, however, inflationary economies—at least the less-developed ones—tend also to be "financially repressed" in that, partly as a consequence of the high rates of inflation themselves, the functioning of financial and capital markets in such economies tends to be severely distorted. McKinnon (1973, chap. 7) further contends that this phenomenon of financial repression should be integrally incorporated into any analysis of stabilization policy. In particular, it is claimed that, as a result, stabilization through an initial increase in the average nominal interest rate paid

This paper is from a chapter of my doctoral dissertation, submitted to Stanford University in August 1974. I gratefully acknowledge my deep indebtedness to my principal adviser, Professor R. I. McKinnon, for his most valuable guidance at all stages of the dissertation research. The most detailed and painstaking comments of an anonymous referee on an earlier, considerably lengthier, version of this paper are also deeply appreciated. I am, however, solely responsible for all views expressed herein.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]

© 1976 by The University of Chicago. All rights reserved.

on money holdings is likely to have significantly more favorable short-run effects on the time path of real output than is stabilization through an initial reduction in the rate of monetary expansion.

In section I we set up the basic model which seeks to incorporate, in a rigorous albeit simplified form, the key structural elements involved in McKinnon's analysis In Section II, the model is employed to determine the short-run, macroeconomic implications of stabilization through the use of the two monetary policy instruments described above Finally, Section III presents some concluding observations

I

The basic "McKinnon thesis" may be briefly adumbrated as follows In many if not most less-developed countries (LDCs), the commercial banking system tends to be the most important source of external financing for productive enterprises Among other things, commercial banks therefore play a strategic intermediary role as a source of financing for the holding of working capital by such enterprises, where the term "working capital" has been defined by McKinnon to include "inventories of goods in process, trade credit, and advances to workers prior to the actual sales" (1973, p 87) Suppose now that the economy in question is initially in inflationary equilibrium, with high rates of nominal money expansion and of inflation, and that the monetary authorities abruptly engineer a discrete reduction in the rate of monetary expansion to a lower but still positive level One might then plausibly argue that some time must necessarily elapse before this induces an appreciable decline in the expected rate of inflation and thereby an increase in the real stock of money in the economy In the meantime, the reduction in the rate of monetary expansion will, through a mechanism that is described more precisely below, have the effect of reducing the flow of real bank credit below what it would otherwise be This constricts the capacity of productive enterprises to hold working capital, and an immediate reduction in real output may result

It is thus the linkage between bank financing and aggregate supply through the working-capital nexus that is employed to explain both the initial downturn in real output that has historically characterized many attempts at price stabilization and the subsequent expansion of real output as inflationary expectations recede and the real stock of money increases Consistent with this analysis, McKinnon has proposed the use of an alternative monetary policy instrument in the initial stages of the stabilization program. rather than reduce the rate of monetary expansion, the monetary authorities should initially raise the average nominal deposit rate paid on money holdings, thereby reducing the excess supply of money by raising the demand for real money balances. Concomitant with this increase in real money balances there will, as shown below, occur an immediate increase in the flow of real bank credit and hence in real output. Subsequently, the monetary authorities can proceed to gradually reduce the rate of monetary expansion as inflationary expectations decline, so that the flow of real bank credit need not decline at any time during the stabilization process and the short-run squeeze on working capital and real output be completely avoided

We turn now to the construction of our formal model Let the total flow of *fixed* capital services that are available at time t be denoted by $\hat{G}(t)$ Suppose, however, that there exists a technical complementarity between each hour of use of fixed capital and the associated requirement of working capital. If the economy is financially repressed, the capacity of entrepreneurs to finance the holding of working capital will be severely limited this will reduce the actual usage of working capital, and hence, as a result of the above-mentioned technical complementarity, the flow of fixed-capital services that is actually utilized, denoted $G(t)$, will be less than the total availability, $\hat{G}(t)$ Such a situation of persistent excess capacity is perfectly compatible with a steady-state equilibrium (which, as will be seen below, is characterized *inter alia* by a constant rate of growth of $G[t]$) if we make the further assumptions that (*a*) technological indivisibilities in the size of each unit of fixed capital render it impossible for each entrepreneur to exactly "tailor" the size of his machines to the availability of working capital, (*b*) the rate of profit that is yielded from the operation of the capital stock is deemed an adequate return for entrepreneurial effort despite the fact that the capital stock is not fully utilized, and (*c*) owing to transactions and information costs, it is not feasible for entrepreneurs economy-wide to associate and decide to restrain further increases in $\hat{G}$ until the growth in the availability of working capital is sufficient to fully utilize all currently available fixed-capital equipment

Let the total utilized flow of capital services per unit time, both fixed and working, be denoted as $K(t)$ this consists of $G(t)$ plus the usage of working capital Let $G/K = \alpha$, which we have assumed to be fixed by technical factors, thus, the ratio of working capital to the total utilized capital flow is simply $(1 - \alpha)$ [1] For simplicity, we shall assume that the aggregate production function of our one-sector economy is of the Harrod-Domar variety:

---

[1] As the referee has pointed out, this procedure of aggregating the flow of working capital with the flow of services from fixed capital encounters certain conceptual difficulties, insofar as "working capital" encompasses a fairly heterogeneous array of inputs. To simplify the formal analysis, therefore, we shall restrict the definition of working capital to include only inputs of raw materials and goods in process, as well as advances to workers prior to actual sales (in other words, we exclude trade credit from the definition, since this may raise some additional intricacies) Suppose we now adopt the fairly innocuous assumption that, at least in the short run, raw materials, goods in process, and labor services are each required in fixed proportion to output and, further, that the real wage rate does not change during the stabilization process (as would be the case if, following the well-known model of Lewis [1954], employers were to be faced with an infinitely elastic supply of labor at a real wage rate that is a fixed markup over the return

$$Y = \sigma K, \tag{1}$$

where Y = final real output It is, of course, not necessary to account separately for the labor input, since this is treated as part of the flow of working capital

The nominal stock of money, M, is divided into two components:

$$M = C + L \tag{2}$$

In the above expression, C is defined as "high-powered" money, assumed issued by the government as transfer payments, part of it is held as currency in active circulation, and the remainder is deposited with commercial banks The quantity L is equal to the total outstanding volume of loans made by the commercial banking system. The proceeds of these loans are in part withdrawn as currency and in part redeposited as "secondary" bank deposits, which justifies equation (2) above For simplicity, we shall assume that the public's desired and actual currency-deposit ratio and the banks' desired and actual reserve-deposit ratio are unchanged through time, so that L/M and C/M are constant let $L/M = q$, so that $C/M = 1 - q$ One of the monetary policy parameters is thus the rate at which the monetary base, C, is expanded letting dots denote time derivatives, if $\dot{C}/C = \mu$, say, the constancy of q also implies that $L/L = M/M = \mu$

It is necessary next to specify the sources of finance for the holding of working capital These are assumed to be threefold (i) replacement of worn-out working capital that is financed out of the internal resources of productive enterprises, (ii) replacement of worn-out working capital that is financed by borrowing from commercial banks, and (iii) net investment in working capital that is financed by bank borrowing For simplicity, it is assumed that fixed-capital assets do not depreciate and that net investment in working capital is financed solely out of bank borrowing However, once the stock of working capital has been so augmented, its subsequent maintenance is financed partly as in (i) and partly as in (ii) It is further assumed that bank credit is used entirely to finance the holding of working capital

Since the total flow of capital services at time t is $K(t)$, the total usage of working capital at that time is $(1 - \alpha)K(t)$ Again, for simplicity, we shall assume that a fixed fraction θ of this usage is replaced through bank

---

to labor in the "subsistence sector" of the economy) Since ours is a one-sector model, the price of inputs of raw materials and of goods in process is the same as that of final output. It thus follows under the stated assumptions that all relative price weights are constant and, hence, that aggregation of working capital with the services of fixed capital is analytically permissible finance, while $(1 - \theta)$ of it is replaced through internal finance. The ratio θ is constrained to lie in the open interval $(0, 1)$ (as is α). Clearly, entrepreneurs have first of all to repay to the commercial banks the loans they previously incurred in order to purchase the fraction θ of the working capital that they have since used up. These repayments are now available for relending by the commercial banks. thus, the extra amount that the banks have to lend in order to maintain the stock of working capital intact is in nominal terms simply equal to $\dot{P}\theta(1 - \alpha)K$, which represents the increased cost of replacement as a result of a rise in the price level P (Since ours is a one-sector model, the price of working capital is the same as the price of final output, provided, as we shall assume, that the real wage rate is always fixed) Letting $\pi = \dot{P}/P$ denote the economy's rate of inflation (an endogenous variable), the rate of "capital accumulation" in our model is then given by the following equation

$$\begin{aligned} K &= \frac{1}{(1 - \alpha)} \left[\frac{\dot{L} - \dot{P}\theta(1 - \alpha)K}{P} \right] \\ &= \frac{1}{(1 - \alpha)} \left[\frac{\dot{L}}{P} - \pi\theta(1 - \alpha)K \right] \\ &= \frac{1}{(1 - \alpha)} \left[\mu q \frac{M}{P} - \pi\theta(1 - \alpha)K \right] \end{aligned} \tag{3}$$

The numerator of the bracketed expression in the first line of equation (3) is the excess of the increase in nominal bank lending over the increase in the nominal cost of replacing the bank-financed component of worn-out working capital This excess is available to finance net investment in working capital This net increase is then multiplied by $1/(1 - \alpha)$ to obtain the net increase in the utilized flow of total capital services, both fixed and working For example, if working capital is one-quarter of K (so that $1 - \alpha$ equals $1/4$), then a unit net increase in the availability of working capital is capable of sustaining a fourfold increase in the total utilization of capital (given the presence of excess capacity in the economy) [2]

---

[2] Certain observations regarding the above derivation are noteworthy (i) The reader will undoubtedly have noticed that no allowance is made in eq (3) for the repayment of loans incurred in previous periods to finance net (as opposed to replacement) investment in those periods The implicit assumption here is that such loans are viewed by productive enterprises as part of their "setup" costs for new projects and are never amortized (repaid) This assumption vastly simplifies the non-steady-state analysis conducted below without, in my opinion, biasing the qualitative results of this paper (ii) Some readers may at first glance consider implausible the assumption that commercial banks finance the whole of net investment in working capital and only a fraction θ of replacement investment in working capital This, however, is entirely consistent with the notion of "financial repression," whereby producers do not have access to external nonbank financing to initiate new projects involving net investment in working capital They are,

A convenient money demand function with which to work is that originated by Cagan (1956) (money being assumed to be held because it is the sole financial asset available to savers):

$$\frac{\hat{M}}{P} = Ye^{-a(\pi^* - d)}, \tag{4}$$

where a is a positive parameter, d is the average nominal interest rate paid on money holdings, π^* is the expected rate of inflation, and $\hat{M}/P$ denotes desired holdings of real money balances (M/P should then be taken to refer to actual holdings of real money balances).

Let $\gamma = \dot{Y}/Y$ denote the economy's growth rate. Dividing equation (3) by K, noting from equation (1) that $Y/K = \sigma$, and hence that $K/K = \dot{Y}/Y = \gamma$, we have

$$\gamma = \mu \frac{\sigma q}{(1 - \alpha)} \frac{M}{PY} - \pi\theta \tag{5}$$

---

however, in a position to undertake some replacement investment out of the sales proceeds from ongoing projects. Implicitly, it is assumed that, if entrepreneurs did have some additional funds of their own, these are either invested in fixed-capital accumulation or have already been "stretched to the limit" in investment projects undertaken in past periods. Clearly, the "story" we are telling would not be very consistent if entrepreneurs had significant uninvested funds of their own which, despite the presence of underutilized fixed capacity, they were holding back from net investment in perfectly divisible working capital until an adequate volume of accompanying bank financing were made available. Nevertheless, it should be pointed out that our analysis can be generalized somewhat to allow for this possibility. Let us assume that a fraction θ_1 of all net investment in working capital and a fraction θ_2 of all replacement investment are financed through bank loans. In that case, eq. (3) becomes

$$K = \frac{1}{\theta_1(1 - \alpha)} \left[\mu q \frac{M}{P} - \pi\theta_2(1 - \alpha)K\right], \tag{3'}$$

since the term within brackets now represents only bank-financed net investment and not total investment in working capital. If θ_1 is sufficiently close to unity, then use of eq. (3′) in place of eq. (3) would not change any of our results. To simplify the notation, however, we shall continue to work with eq. (3). If θ_1 is significantly less than unity, then the results of this paper would certainly be altered. However, for the reasons just indicated, the notion of "financial repression," which we are seeking to model in this paper, would become quite nebulous were producers not significantly dependent on bank financing to initiate new projects. (iii) As the referee has pointed out, owing to the presence of underutilized fixed capacity in our model, the concept of "capital accumulation" has to be *interpreted not as the augmentation of the total stock of capital in the economy* but, rather, as the augmentation of the stock of working capital coupled with the associated increase in the rate of utilization of fixed capacity. The fact that eq. (3) specifies a dependence of K on L/P suggests that it might be possible to reformulate the model in a more standard fashion by treating money as a "producer's good." This, however, is misleading, as would be immediately evident if we were to contemplate the nature of this functional dependence. If money were a producer's good, then the manner in which the level of M/P influences the level of Y would depend on such considerations as the efficacy with which the use of money frees resources that would otherwise be engaged in exchange for employment in productive activity. Such considerations are, however, wholly absent from the argument in this paper leading to the derivation of eq. (3) above. This simply illustrates the fact that, conceptually, the role of bank lending in our model is primarily to facilitate investment (in working capital). Its effect on the level of production in the economy derives from its effect on investment instead of constituting a separate structural datum in its own right.

Also, logarithmically differentiating the identity $MV = PY$ with respect to time (V being the velocity of circulation of money) and substituting terms, we have

$$\mu + \frac{\dot{V}}{V} = \pi + \gamma \tag{6}$$

Before turning to the analysis of short-run dynamics, let us briefly examine the steady-state properties of the system corresponding to given initial values of the policy parameters μ and d The steady state in our model is defined by the following conditions (i) that actual holdings of real cash balances are equal to desired holdings, the latter being given by equation (4) (outside the steady state, we shall allow for the possibility of a discrepancy between these two variables), and (ii) that the expected rate of inflation π^* is equal to the actual rate π and that both of these are constant The latter implies that velocity is constant, so that equation (6) specializes to $\mu = \pi + \gamma$ From this, we have $\pi^* = \pi = \mu - \gamma$, substituting this into equation (4) and then substituting equation (4) into equation (5), we have in the steady state

$$\gamma = \mu \frac{\sigma q}{(1 - \alpha)} e^{-a(\mu - \gamma - d)} - \mu\theta + \gamma\theta \tag{7}$$

This is an implicit equation in the single unknown γ Rewriting it as $\gamma - \mu[\sigma q/(1 - \alpha)]e^{-a(\mu-\gamma-d)} + \mu\theta - \gamma\theta = 0$, it can be shown that, under some relatively weak restrictions, this equation has in general two solutions for each combination of values of μ and d, at one of which its "Jacobian" is positive and at the other negative Given the dynamic adjustment mechanisms that we postulate below, however, it can be shown that only the former solution satisfies the necessary condition for the local stability of the model We shall therefore restrict our attention to this category of solutions For these, it can further be demonstrated that there exists a unique, positive value of μ which maximizes the resulting value of γ (holding d fixed) As μ is increased above this unique value, the corresponding value of γ falls progressively to zero (and below) We shall assume that initially the economy lies in this latter, "financially repressed" region, so that a reduction in μ raises the steady-state solution for γ (and hence, a fortiori, lowers the steady-state value of π)

We turn now to the analysis of short-run dynamics First of all, as regards the evolution of π^* through time, we shall simply adopt Cagan's adaptive expectations formulation,

$$\frac{d\pi^*}{dt} = \beta(\pi - \pi^*), \tag{8}$$

where β is a positive parameter

The purpose of our final specification, to which we now turn, is to incorporate into the model an important facet of empirically observed stabilization experience This is the tendency for the velocity of circulation V to initially rise in response to a discrete reduction in μ, despite the fact that π commences falling almost immediately. It is only after some further time has elapsed that V begins to fall and to eventually decline to a new, lower, steady-state value Within the context of our present analysis, this initial phase of a rising velocity assumes particular significance, since McKinnon has argued (1973, pp 84–85), in his interpretation of the Brazilian stabilization experience of the mid-1960s, that this would constitute a phase of increasing "disintermediation" of the banking system (since a rising V implies a falling money-GNP ratio) and, hence, increasing stringency in the availability of loan-financed working capital

McKinnon has also suggested (private communication) that this initial rise in velocity might be generated if we were to assume that the actual rate of inflation is a function, *inter alia*, of the expected rate of inflation Indeed, models which assume that π is a function of π^* and of "demand pressure" have been proposed by Stein (1970), Hadjimichalakis (1971), Goldman (1972), and Fischer (1972) (The last-mentioned author has also suggested that a possible microeconomic rationale for this specification would be that sellers raise output prices in part in response to expected increases in input costs None of these authors, however, considers the implications of this specification for the short-run behavior of velocity) Analogously to Goldman, we may write

$$\pi = f\left(\frac{\text{excess demand for goods}}{\text{goods supply}}\right) + \pi^*, \tag{9}$$

where f is an increasing and sign-preserving function of demand pressure By Walras's law, however, excess demand for goods must equal the excess supply of money, since these are the only two markets in the model Thus, we can change equation (9) to

$$\pi = h\left(\frac{M}{PY} - \frac{\hat{M}}{PY}\right) + \pi^*, \tag{10}$$

where for simplicity we shall take h to be a positive constant (incorporating, *inter alia,* some multiplicative factor k, say, representing the fraction of the stock disequilibrium between actual and desired money holdings which money holders choose to eliminate at any given time) and where M/P and $\hat{M}/P$ denote actual and desired holdings of real cash balances, respectively

This completes our specification of the model To facilitate the mathematical analysis, it is convenient to introduce some additional notation. Let V equal the actual level of velocity, which in turn is equal to PY/M, and let $\hat{V}$ be the desired level of velocity or $PY/\hat{M}$ Moreover, let W = log V and $\hat{W}$ = log $\hat{V}$ (all logarithms are to base e) Equation (6) then becomes

$$\mu + \dot{W} = \pi + \gamma \tag{6a}$$

Moreover, $M/PY = 1/V = e^{-W}$, since $\log 1/V = -W$. Equation (5) can therefore be written as

$$y = \mu \frac{\sigma q}{(1 - \alpha)} e^{-W} - \pi\theta. \tag{5a}$$

Similarly, $\hat{M}/PY = e^{-\hat{W}}$, and equation (10) becomes

$$\begin{aligned} \pi &= h(e^{-W} - e^{-\hat{W}}) + \pi^* \\ &= h[e^{-W} - e^{-a(\pi^* - d)}] + \pi^*, \end{aligned} \tag{10a}$$

since, from equation (4), $\hat{W} = \log PY/\hat{M} = a(\pi^* - d)$

Substituting from equations (5a) and (10a) into equation (6a), we can express $\dot{W}$ as a function of W and π^* Similarly, substituting from equation (10a) into the adaptive expectations function, which is equation (8) above, we can express $d\pi^*/dt$ as a function of these same variables The model is thereby reduced to the following self-contained dynamic system

$$\begin{aligned} \dot{W} = &- \mu\left(1 - \frac{\sigma q}{1 - \alpha} e^{-W}\right) + (1 - \theta)\pi^* \\ &+ (1 - \theta)h[e^{-W} - e^{-a(\pi^* - d)}], \end{aligned} \tag{11}$$

$$\frac{d\pi^*}{dt} = \beta h[e^{-W} - e^{-a(\pi^* - d)}] \tag{12}$$

The steady state in this model is defined by the conditions that $\dot{W} = 0 = d\pi^*/dt$ These can easily be shown to imply that $\pi = \pi^*$ and $W = \hat{W} = a(\pi^* - d)$ In what follows, we assume that the dynamic system is stable [3]

A diagrammatic representation of this system is illuminating From equations (11) and (12), it can be readily verified that the condition that the Jacobian be positive implies that the locus of combinations of W and π^* along which $\dot{W} = 0$ must, at least in some open neighborhood of the steady-state equilibrium, be steeper than the locus along which $d\pi^*/dt = 0$ We therefore have the phase diagram, in W-π^* space, shown in figure 1

It is not difficult to verify from equations (11) and (12) that the vectors in figure 1 have the directions indicated Starting from any arbitrary initial position, such as B, the system will as drawn converge to the steady-state equilibrium, point A In the figure, the convergence is depicted as being cyclical in character This need not, of course, always be the case.

---

[3] Linearizing eqq (11) and (12) about the steady-state equilibrium, we know that a necessary and sufficient condition for the local stability of this dynamic system is that the trace of the resulting matrix be negative and its determinant positive, where both of these are to be evaluated at the point $W = a(\pi^* - d)$ The trace is equal to $[-\mu(\sigma q)/(1 - \alpha) - (1 - \theta - a\beta)h]e^{-W}$ We shall assume that this entire expression is negative. The determinant is simply equal to the Jacobian of eq (7) above, thereby establishing our earlier assertion that only that steady-state solution for y at which the Jacobian is positive can be locally stable.

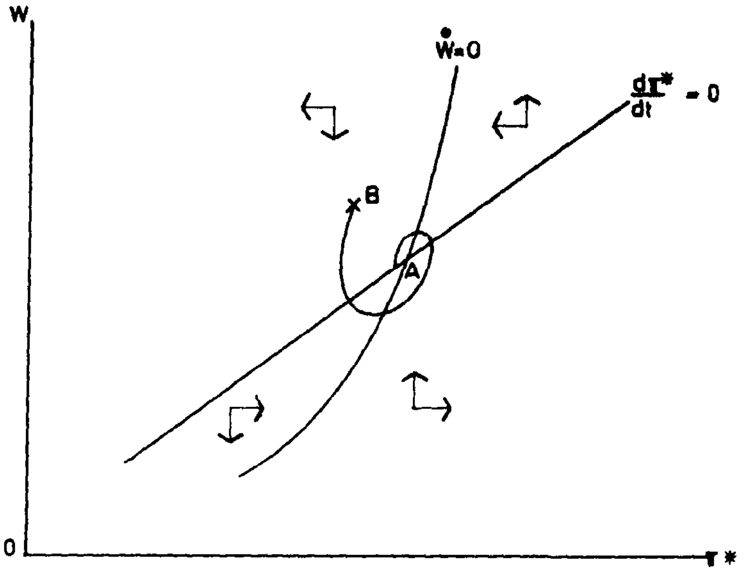


FIG 1

II

We can now employ the apparatus just developed to examine the short-run implications of alternative stabilization policies The first policy to be discussed is a sudden reduction by the monetary authorities of the rate of monetary expansion Suppose that μ is initially so high that the steady-state value of γ is equal to zero[4] and that the monetary authorities institute a discrete reduction in μ to a somewhat lower but still positive level, as a result of which the new steady-state value of γ becomes positive. What about the transitional response of W, π^*, π, and γ? From equation (12), it is seen that the reduction in μ has no effect on the locus $d\pi^*/dt = 0$ in figure 1 As regards the effect on the locus $\dot{W} = 0$, it can be readily shown from equation (11) that, if the term $1 - [\sigma q/(1 - \alpha)]e^{-W}$, evaluated at point A in figure 1, is strictly positive (we show below that this condition will always hold), then this locus, being continuous, will shift upward in at least some open neighborhood of the point A. The transition to the new steady-state equilibrium may therefore be depicted as in figure 2.

Clearly, the vectors in figure 2 have to be "oriented" toward the new steady-state equilibrium, which is point C rather than the previous

---

[4] In economic terms, a zero value of γ simply implies that financial repression has become so pronounced that the entire increase in nominal bank credit is always used for replacement purposes

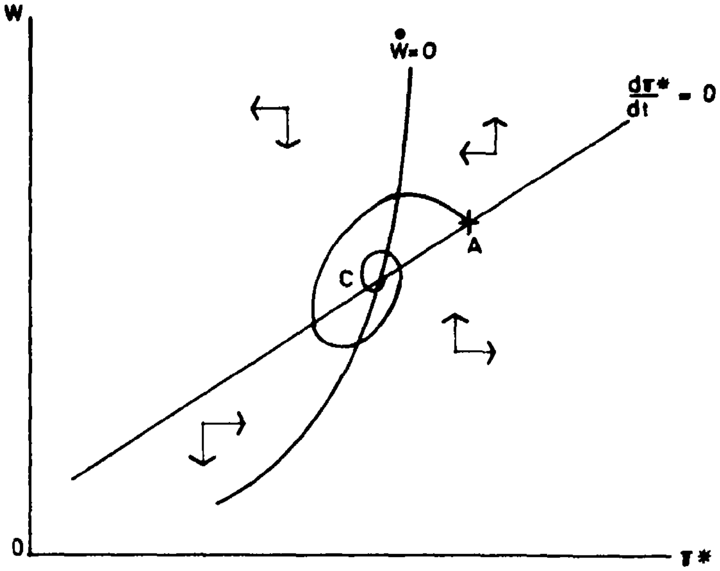


FIG 2

equilibrium, point A Thus, starting at, say, time 0 at point A, the effect of reducing μ will be to induce first of all a phase of rising W (and hence rising velocity) and falling π^*, followed by a phase of falling W and falling π^*, followed by a possibly cyclical convergence to the new equilibrium point Once again, however, the convergence need not necessarily be cyclical in character

Before explicating the economic rationale for these time paths, let us consider the impact of a reduction in μ on the other two variables of interest, π and y From equation (10a) we see that, since W and π^* are continuous functions of time, the downward shift in μ does not generate any instantaneous movement or jump in π, in other words, at time 0, π is completely unaffected by the reduction in μ Immediately thereafter, however, the rise in W (which implies a reduction in the real supply of money relative to output) and fall in π^* (which increases the real demand for money relative to output) jointly generate an excess demand for money, which causes π to fall (this effect is in addition to the fact that a falling π^* itself causes π to fall) Thus, in the first phase delineated above, π is unambiguously falling In the second phase, π^* is still falling, which would tend to cause π to continue falling However, in this phase W is also falling, implying an augmentation in the real supply of money relative to output, and this could very well offset the effect of a falling π^* and cause π to rise instead

In order to resolve this and other indeterminacies discussed below, some numerical simulations of the model were undertaken, the details of which are available from the author upon request. The simulations indicate that even in this second phase, π will continue falling; evidently, the effect of the falling π^* outweighs that of the falling W. In fact, π continues falling until it "overshoots" its new, lower, steady-state value, after which it gradually converges upward to this new value

The effect on y is given by equation (5a) above. At time 0, neither W nor π jumps in response to the reduction in μ; therefore, the reduction in μ causes an instantaneous downward jump in y, and since y was assumed initially to equal zero, this implies that y initially turns negative, so that real output commences declining [5] The economic rationale for this result is that the reduction in μ instantaneously reduces the *net* flow of real bank credit (in other words, the flow of bank credit in excess of the lending out of repaid loans) However, the volume of bank credit required to finance replacement investment in working capital is growing at rate π, which has yet to adjust downward Thus, the flow of bank credit is insufficient to meet the replacement demand, leading to negative net investment in working capital and a fall in real output

After time 0, y is subject to opposing influences On the one hand, in the first phase, π is falling, which would tend to cause y to rise On the other hand, W is rising, which further reduces the net flow of real bank credit (this time owing to "disintermediation"—a reduction in the real size of the banking system relative to output), which would tend to cause y to fall further Moreover, our simulations suggest that, for what appear to be equally "reasonable" alternative parameter specifications, either possibility can in fact occur However, the simulations also indicate that even when y continues falling after time 0, this decline will reverse itself shortly thereafter, while W is still rising [6] In other words, the falling π has the effect, early in the process, of overwhelming the continuing disintermediation of the banking system to generate an upturn in y

In the second phase, W is falling, so that the real size of the banking system relative to output, and hence the net flow of real bank credit, is rising. In addition, the simulations indicate, as pointed out above, that π will be falling in this phase On both counts, therefore, y will continue rising and will eventually overshoot its new, higher, steady-state value and then converge downward to it.[7]

---

[5] If y were initially positive, the reduction in μ would still generate an initial fall in y However, depending on what its initial value as well as the magnitude of the reduction in μ were, y may or may not initially turn negative.

[6] The fact that some such phenomenon must occur can be established from continuity considerations alone That is to say, since W is a continuous function of time after time 0 and since along the path $\dot{W}$ equals zero when the path crosses the $\dot{W} = 0$ locus while $d\pi/dt$ can be shown to be strictly negative (and continuous) at that point, it follows that from at least some time prior to that point onward, the effect of the falling π on y must outweigh that of the rising W (since y is continuous in π and W)

Let us now temporarily retrace our steps to examine why velocity initially rises in our model, although an explanation of this will have to be somewhat heuristic in character, since velocity is both influenced by and influences the other variables of the model Fundamentally, the rise in velocity is attributable to the fact that π does not adjust downward as rapidly as μ. That is to say, velocity V is equal to PY/M, initially, P/M must rise as P rises as fast as in the previous equilibrium, while M rises at a slower rate It is true that this effect is somewhat offset by the falling Y. However, from equations (6a) and (5a), it is easily shown that, provided $[\sigma q/(1 - \alpha)]e^{-W} < 1$ at time 0,[8] Y will not fall fast enough to offset completely the rising P/M It is only subsequently, once π has fallen sufficiently (under the influences of falling expectations and increasing excess demand for money), that velocity will cease rising and commence falling

We turn next to a consideration of the second stabilization policy instrument, the nominal deposit rate d Let l be the nominal interest rate charged on bank loans If we assume that (i) banks are in competitive equilibrium with zero excess profits, (ii) the government pays no interest on bank reserves or on currency in circulation, and (iii) the costs of the monetary system are a constant fraction z of the real money supply M/P, then it is easily shown that $d = ql - z$ Hence, barring governmental subsidies to the banking system, a change in d requires a concomitant change in l. Although we shall henceforth be referring to changes in d, the corresponding changes in l will always be assumed to have been effected.

---

[7] It should be pointed out that, while the reduction in μ (and also, as will be seen below, an increase in d) is capable of producing a movement toward a higher growth rate of the economy, the persistence of this higher growth rate over time presupposes the existence of some mechanism which ensures that the availability of fixed-capital services, $\hat{G}(t)$, expands at an adequate rate, since otherwise at some point the fixed capacity would become fully utilized Since our concern in this paper is primarily with short-run analysis, we shall not pursue this matter further, except to point out that an obvious candidate for such a mechanism is an increase in fixed investment out of the higher level of business profits that results from a higher level of capacity utilization.

[8] This inequality states that the real volume of bank loans in the original steady-state equilibrium is strictly less than the real stock of working capital in the economy. If γ is initially zero, then this condition is trivially satisfied, since the real volume of bank loans would simply equal the fraction θ of the real stock of working capital If γ were initially positive, then, by integrating eq (3) along the initial steady-state path, it can be shown that this condition will be satisfied if and only if $\theta\pi < \pi$, which will hold as long as π is originally positive Essentially, in this latter case, commercial banks allocate some loans at each point in time to financing net investment in working capital In subsequent periods, as prices rise the real value of these past loans would fall However, the real value of the working capital that has thereby been "brought into existence" remains unchanged, and the fraction $(1 - \theta)$ of this is continually being replaced from the internal resources of productive enterprises In steady-state equilibrium, therefore, the real volume of bank loans will fall short of the real stock of working capital Intuitively, it is precisely because businesses are not totally dependent on bank credit to finance their holdings of working capital that γ at time 0 does not jump downward by as much as μ (the fall in μ being in absolute value equal to the percentage rate of increase of P/M at time 0)

By implicitly differentiating equation (7) with respect to d, it is readily verified that an upward shift in d raises the steady-state value of γ, which is what we would expect, since it raises the real size of the banking system and hence the net flow of real bank credit to finance investment in working capital. This raises the question as to whether there is any upper bound to the feasible range of values that d may assume, a question which is taken up below. As regards short-run dynamics, an upward shift in d will, at least in some open neighborhood of the point A in figure 1, cause both of the loci in that figure to shift downward. However, from equations (11) and (12), it can be easily shown that the locus $d\pi^*/dt = 0$ will definitely shift down by more, so that the new steady-state value of π^* is lower (which is consistent with the fact that the new steady-state value of γ is higher, while μ is unchanged). The dynamic evolution of the system to its new equilibrium is thus as shown in figure 3.

The new steady-state equilibrium is given by point C' in figure 3. We thus see that, unlike the case of a reduction in μ, an upward shift in d generates an initial phase of falling W (and hence falling velocity) and falling π^*, both of which then proceed to converge in a possibly cyclical manner to C'.

Before providing an intuitive explanation for this result, let us examine the time paths of π and γ. At time 0 itself, there is no jump in either W or π^*; thus, from equation (10a), we see that the upward shift in d produces an instantaneous downward shift in π at time 0. This is attributable to the fact that the increase in d immediately increases the desired level of real money holdings (although the actual level does not change at that instant) and thereby creates an instantaneous excess demand for money. An increase in d therefore exerts a more rapid deflationary effect than a reduction in μ (since, as pointed out earlier, the latter has no effect on the value of π at time 0). As McKinnon has suggested to me, this is due to the fact that an increase in d directly increases the expected real return to money perceived by money holders, whereas a reduction in μ can only increase the expected real return after it has worked through the entire system to produce a fall in π^*.[9]

---

[9] Following on this discussion, the referee has suggested that "the comparison between a μ-policy and a d-policy—which in some sense is the core of the paper—is conditional upon the sluggishness inherent in the adaptive expectations hypothesis." I must confess that this statement contains an important element of truth. If, that is to say, individuals were to link their expectations of future inflation rates directly to movements in μ, the deflationary impact of a reduction in μ would certainly be accelerated. (There would still remain some differences in the response of the system as compared with the effect of an upward shift in d, however, in view of the fact that the latter does not, while a reduc-

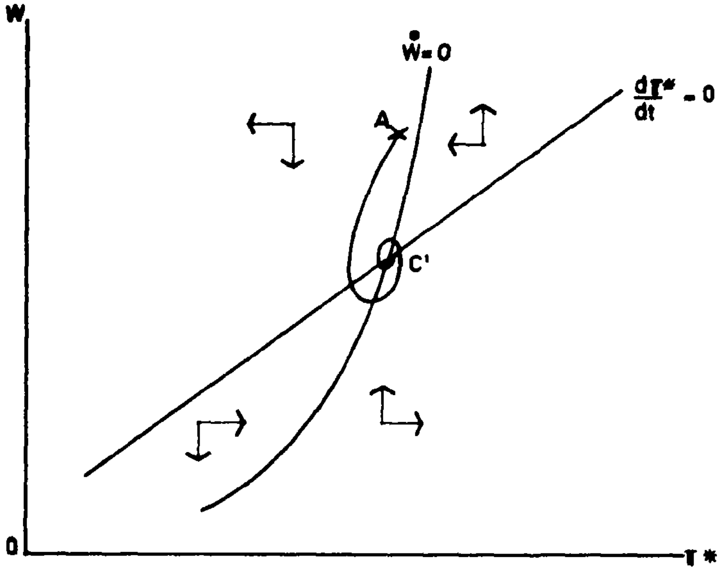


FIG 3

After time 0, π is subject to opposing influences On the one hand, π^* commences declining, which would tend to cause π to fall further. On the other hand, *W* also declines, thereby augmenting the real supply of money relative to output, which would tend to cause π to rise Our simulations indicate, however, that the former effect dominates, so that π continues declining The simulations also indicate that π will decline until it overshoots its new, lower, equilibrium value and then converges upward to it

As regards the time path of γ, from equation (5a) it is seen that, since, at time 0, π registers an instantaneous downward jump, γ registers an instantaneous upward jump. This occurs because the increase in *d* has no instantaneous effect on the net flow of real bank credit However, by reducing π, it reduces the rate of increase in the cost of replacing worn-out

---

tion in μ does, entail a simultaneous reduction in the net flow of real bank credit Consequently, as may be seen from the following discussion, one may expect the effect on γ to continue to be more favorable in the case of an increase in *d*) Owing to prevailing low levels of literacy and economic sophistication in LDCs, however, there is strong reason to believe that the adaptive expectations hypothesis remains applicable in this context. This is particularly so in view of the fact that, since the steady-state value of γ is an endogenous variable in our model, movements in μ do not induce exactly equal movements in the steady-state value of π—unlike the case of "developed-economy" models, in which γ is exogenously fixed I am convinced, therefore, that the present model remains prescriptively useful at least as far as LDCs are concerned.

working capital Some portion of the net flow of real bank credit is thus released to finance net as opposed to replacement investment in working capital, thereby generating an upward shift in γ After time 0, W commences falling This implies an increase in the real size of the banking system relative to output, and hence an increase in the net flow of real bank credit This would tend to further increase γ. Moreover, since the simulations suggest that π continues falling after time 0, γ will, on both counts, continue rising, until it rises above its new, higher equilibrium value, and thereafter converges downward to it [10]

From the adaptive expectations function given in equation (8), it is clear that, since π registers a downward jump, π^* must commence falling for at least some time after time 0 The rationale for the fall in velocity is slightly more subtle V is equal to PY/M, and since M grows at the same rate as before while P grows at a lower rate, P/M must commence falling This is to some extent offset by the rise in Y However, from equation (5a) it is seen that, since $\theta < 1$, the upward shift in γ is lower than the downward shift in π,[11] so that on balance velocity must commence falling

It is seen, therefore, that an upward shift in d has favorable effects on π and γ in both the short run and the steady state This raises the question as to whether there is any upper bound on the magnitude of permissible increases in d Such a bound does exist by virtue of the fact that, as pointed out earlier, changes in d require concomitant changes in l, the nominal loan rate It is clear that both in the steady state and at each point along the economy's transition path, the expected real loan rate $l - \pi^*$ should not be allowed to exceed the real rate of profit on working capital, since otherwise business borrowing from commercial banks for the purpose of investing in working capital will fall to zero (notice that π^* varies in value during the transition from one steady-state equilibrium to another). Since working capital is equal to $(1 - \alpha)K$, the real rate of profit on working capital is equal to $[Y - (1 - \alpha)K]/(1 - \alpha)K$, which in turn is equal to $[\sigma - (1 - \alpha)]/(1 - \alpha)$, which we shall denote as $\hat{r}$. Since $d = ql - z$, the constraint $l - \pi^* < \hat{r}$ is equivalent to the condition $d < q(\hat{r} + \pi^*) - z$, the latter therefore constituting our (variable) upper bound. As a matter of fact, in view of the observation made in footnote 7 above, one would ideally require that d be fairly appreciably less than this upper bound in order to provide businesses with some net profits from which to finance fixed-capital investment. However, we shall not attempt to formally integrate investment in fixed and in working capital in this paper.

---

[10] The reader may at this point be tempted to inquire into the "robustness" of our results, given that we have found it necessary to resort to numerical simulations to resolve certain indeterminacies in the analysis Analytically, however, it is possible to definitely establish two conclusions First, since at time 0, γ registers an instantaneous upward jump when d is raised and an instantaneous downward jump when μ is reduced, and since the time path of γ is a continuous function of time after time 0, both at time 0 and for at least some time thereafter the time path of γ resulting from an increase in d will dominate (be higher than) the time path of γ resulting from a reduction in μ Second, by similar reasoning, the time path of π resulting from an increase in d must, at least in the early time periods, be lower than that resulting from a reduction in μ These considerations do establish a presumption in favor of stabilization through an initial increase in d A complete normative evaluation is, however, postponed until the following section.

[11] This is attributable to the fact that a reduction in π affects γ by releasing some bank credit from replacement financing to finance net investment in working capital. Thus, since $\theta < 1$ (so that not all replacement investment is financed by the commercial banks), each 1 percent increase in the volume of bank-financed holdings of working capital (resulting from a one percentage point decrease in the inflation rate) implies a less than 1 percent increase in the volume of total holdings of working capital, and hence a less than one percentage point increase in the economy's growth rate.

III

This paper has sought to fulfill a twofold purpose. In the first place, we have attempted to specify explicitly a logically self-contained set of analytical relationships consistent with McKinnon's analysis of stabilization experience in financially repressed economies. The relationship among the flows of bank credit, working capital, and real output has been formally characterized.[12] Moreover, we have demonstrated the importance of "building into" the model an "inflationary momentum" in the form of a dependence of the actual rate of inflation on the expected rate as well as on demand pressure, in order to account for the characteristic short-run behavior of real output and of the velocity of circulation of money that typically results from a reduction in the rate of monetary expansion. In the second place, we have shown, within the context of our formal model, that McKinnon's "inside money" approach to stabilization policy appears to be valid and fruitful: price stabilization through an initial increase in the nominal deposit rate paid on money holdings enables the economy to experience an immediate accompanying gain in real output, instead of the fall that is typically attendant on a reduction in the rate of monetary expansion.

---

[12] This formal specification does, incidentally, enable us to clear up a confusion between stocks and flows that appears to have crept into McKinnon's analysis. He has suggested (1973, pp. 86–87) that the reason that a reduction in the rate of monetary expansion precipitates a temporary fall in the level of real output is that it causes an initial contraction in the real stock of money and hence in the real stock of bank credit, which we have denoted as qM/P. This appears to be a slightly misplaced emphasis. What our model suggests is that real output falls because the real flow of *net* bank credit, which is equal to $\mu qM/P$, falls initially. The latter is only in part a function of the real stock, qM/P. In our model, e.g., at time 0, M/P is unchanged, but real output begins to fall because $\mu qM/P$ falls. Moreover, it should be emphasized that the evolution of the economy's growth rate over time is a function not solely of the movement in $\mu qM/P$ but, rather, of the movement in $\mu qM/P$ *in relation to* the movement in the rate of inflation. Thus, in some of our simulations, M/P commences falling after time 0, and yet the economy's growth rate commences rising because π commences falling even faster.

From the foregoing, it would appear that the "optimal" means of stabilization is through an increase in the nominal deposit rate, holding the rate of monetary expansion fixed This, however, would be an illegitimate extrapolation of our results. For one thing, as pointed out earlier, the nominal deposit rate cannot be raised beyond the point where the real loan rate equals the rate of profit on working capital. More fundamentally, however, it is very likely that the monetary authorities are desirous of achieving not only a target rate of growth but also a target rate of inflation Since we thus have two policy objectives, it follows from the standard Tinbergen analysis that both instruments, the nominal deposit rate and the rate of monetary expansion, will have to be actively deployed[13] What our analysis strongly suggests, however, is that initially the nominal deposit rate should be raised Once this has induced a decline in the inflation rate (and an increase in the growth rate), the other instrument can be brought judiciously into play Ideally, of course, it would be desirable to specify target values for the inflation rate and the growth rate, and a loss function, and then to choose the values of d and μ over time in such a manner as to minimize the discounted integral of the loss function over a finite or infinite time horizon Such an optimization exercise (which would have to be subject to the above-mentioned constraint on d) is not attempted here, however, the analysis of this paper does hopefully provide many of the ingredients required for such an exercise

References

Cagan, P "The Monetary Dynamics of Hyper-Inflation " In *Studies in the Quantity Theory of Money*, edited by M Friedman Chicago Univ Chicago Press, 1956

Fischer, S "Keynes-Wicksell and Neoclassical Models of Money and Growth " *A E.R* 62, no 4 (December 1972) 880–90

Frenkel, J. "Inflation and the Formation of Expectations " *J Monetary Econ.* 1 (October 1975) 403–21

Friedman, M "A Theoretical Framework for Monetary Analysis " *J P E* 78, no 2 (March/April 1970) 193–238

Goldman, S M "Hyperinflation and the Rate of Growth in the Money Supply " *J Econ Theory* 5 (October 1972) 250–57.

Hadjimichalakis, M G "Money, Expectations, and Dynamics An Alternative View " *Internat Econ Rev* 12, no 3 (October 1971) 381–402.

Lewis, W A "Economic Development with Unlimited Supplies of Labour." *Manchester School Econ and Soc Studies* 22, no 2 (May 1954): 139–91.

McKinnon, R I *Money and Capital in Economic Development* Washington Brookings Inst., 1973.

---

[13] It can be shown that the two instruments are "independent" in the sense that the set of values of the target variables obtainable by manipulating one instrument differs from that obtainable by manipulating the other

Mundell, R. A. "Growth, Stability, and Inflationary Finance" *J.P.E.* 73, no. 2 (April 1965): 97–109.

Selowsky, M. "Cost of Price Stabilization in an Inflationary Economy." *Q.J.E.* 87, no. 1 (February 1973): 44–59

Shaw, E. S. *Financial Deepening in Economic Development.* Oxford. Oxford Univ Press, 1973.

Stein, J. L. "Monetary Growth Theory in Perspective." *A E.R* 60, no. 1 (March 1970): 85–106

Direct versus Indirect Remedies for Externalities

Jerry Green

Harvard University

Eytan Sheshinski

Hebrew University of Jerusalem, Jerusalem, and Stanford University

This paper is concerned with tax policies designed to obtain an improved competitive allocation in the presence of consumption externalities. It is known that the full optimum can, in general, be attained only through the imposition of excise taxes at different levels for different individuals Since these may be ruled out (possibly because of implementation costs), one is confined to consider second-best taxes The common interpretation of the Pigouvian principle has called for taxes on the externality-creating commodities With no relationships between the consumption of different commodities the Pigouvian principle is obviously impeccable. But the existence of substitutes or complements for an externality-causing commodity raises the possibility of indirect policies treating the externality through the markets for related goods Obviously, if the direct policy is not feasible, the indirect treatment may provide some partial remedy. We show, however, that even when direct policies are available, the overall optimum may involve only indirect policies An example with such a result is provided in the paper We also list a number of cases in which the traditional prescription is confirmed, and the overall optimum involves only direct policies

The economic analysis of situations involving externalities has focused on two related issues The first is the comparison between competitive equilibrium and Pareto-optimal allocations, while the second concerns tax policies designed to obtain an improved competitive allocation. The degree of improvement depends, of course, on the scope of the

The work of the first author was supported by NSF grant GS-31688 at Harvard University, and of the second author by NSF grants GS-35890 at the University of California at Berkeley, and SOC-74-11446 at Stanford University

[*Journal of Political Economy*, 1976, vol 84, no 4, pt 1]

© 1976 by The University of Chicago All rights reserved

available fiscal instruments and on the form of the externalities. It is known that the full optimum can, in general, be attained only through the imposition of excise taxes at different levels for different individuals. As we shall show below, a notable exception to this rule is the case of consumption externalities which depend on the total consumption of certain commodities by the economy provided that individuals disregard the effect of their actions on the level of externalities In this case, which Meade (1952) appropriately termed "atmosphere" externalities, ordinary commodity taxes are sufficient for achievement of the full optimum. This type of externality is the analog of the perfectly competitive model, because in the large-numbers case the individual would be correct in neglecting the influence of his consumption on the level of the externality in the aggregate In more general cases, for example, with a small number of consumers or with large numbers but asymmetries in individuals' behavior or in externality generation, policies discriminating between individuals are required for attaining a full Pareto optimum Since these may be ruled out (possibly because of implementation costs), one is confined to consider second-best taxes This is the case treated in this paper

The problem of the optimal tax policy is to determine the commodities upon which taxes are imposed (or subsidies granted) and the optimum levels of these taxes The Pigouvian principle (Pigou 1932, pp 192 ff), accepted by all writers (see, for example, Baumol [1972] and the references therein), has been to impose a tax/subsidy on the generator of the externality In the case of nonreciprocal negative production externalities (smoke affecting costs of neighboring laundries), this is achieved by imposition of effluent charges aimed at restricting output of the externality-creating factor In the case of reciprocal consumption externalities, the common interpretation of the Pigouvian principle calls for taxes on the externality-creating commodities Again, Davis and Whinston (1962), Plott (1966), Buchanan (1969), and others have shown that according to this principle anomalous cases cannot be ruled out For example, with negative externalities an improvement may be realized by granting a subsidy rather than imposing a tax

This literature has completely neglected the question of the proper variables to be taxed (or subsidized) With no relationships between the consumption of different commodities the Pigouvian principle is obviously impeccable But the existence of substitutes or complements for an externality-causing commodity raises the possibility of indirect policies· treating the externality through the markets for related goods Thus, if cars create smog, one remedy may be to tax goods complementary with car trips (such as recreation).

It seems plausible that indirect policies are in general inferior to direct policies. In the international trade literature, for example, one finds statements that in the presence of market distortions (which may be viewed as a kind of externality) direct fiscal policy in these markets is always superior to indirect policies (Johnson 1972).

On the basis of a more general model, we reach a different conclusion. Obviously, if the direct policy is not feasible, the indirect treatment may provide some partial remedy We show, however, that even when direct policies are available, the overall optimum may involve only indirect policies. An example with such a result is provided below Diamond (1973) has a similar example where the direct policy fails, but he has not considered the possibility of indirect policies We also list a number of cases in which the traditional prescription is confirmed and the overall optimum involves only direct policies

The Model

We consider a model which has two consumers who purchase an externality-causing good α and a related good β, among other commodities To rule out income effects we assume that utility is linear in income net of expenditures on α and β The unit costs of α and β are p and q Policies considered in this paper are per unit taxes/subsidies on consumption of α and β, denoted t and z, respectively Consumers face prices $p + t$ and $q + z$ for these goods, and maximize their utility functions given by

$$U = u(\alpha_1, \beta_1, \alpha_2) + X_1 + T_1, \tag{1}$$

$$V = v(\alpha_2, \beta_2, \alpha_1) + X_2 + T_2, \tag{2}$$

where $X_i = I_i - (p + t)\alpha_i - (q + z)\beta_i$, $i = 1, 2$, I_i are exogenous incomes of the two individuals, and T_i are lump-sum transfer payments, representing a redistribution of tax revenues

Note that in this formulation the individual's consumption choices of α and β are independent of his money income and his lump-sum taxes, T_i

The transfer payments are thought of as proportional redistributions of the tax revenue If θ is the share of individual 1 in the taxes collected,

$$T_1 = \theta[t(\alpha_1 + \alpha_2) + z(\beta_1 + \beta_2)], \tag{3}$$

$$T_2 = (1 - \theta)[t(\alpha_1 + \alpha_2) + z(\beta_1 + \beta_2)] \tag{4}$$

The first-order conditions for the individuals' maximization problems, assuming that the others' actions are fixed, are

$$\begin{aligned} u_1 &= p + t, & v_1 &= p + t, \\ u_2 &= q + z, & v_2 &= q + z \end{aligned} \tag{5}$$

It is clear that competitive equilibria with $t = z = 0$ are generally inefficient. The goals of this study are to evaluate the comparative efficacy of t and z as policy tools and to investigate the seemingly intuitive presumption that direct corrective taxation is a superior tool.

Negative Result

In this section we present an example showing that indirect taxation may be a useful policy while direct measures are impotent. To do this we demonstrate that $t^* = 0$ is the optimal direct tax when $z = 0$ and that $z^* \neq 0$ is the optimal indirect tax when $t = 0$. A fortiori, the resulting allocation, when indirect taxes are used, is superior.

Consider the case of a direct tax, and let

$$U = \mathcal{U}(\theta, t),$$
$$V = \mathcal{V}(\theta, t), \tag{6}$$

be the realized values of utility in an equilibrium in which t and θ are the policy parameters. These functions can be depicted as in figure 1.

The dashed curves are iso-$\mathcal{U}$ loci in the (t, θ) space. A positive slope indicates $d\mathcal{U}/dt < 0$, since $d\mathcal{U}/d\theta$ is always positive. Restricting ourselves to the policy variables θ and t, a constrained Pareto optimum is characterized as a maximum of $\mathcal{U} + \mathcal{V}$, with respect to these variables. Since there are no income effects, we have $(d\mathcal{U}/d\theta) \equiv -(d\mathcal{V}/d\theta)$.

This explains why all such constrained optima (θ^*, t^*) are determined by t^* alone, since any θ would satisfy the first-order conditions. The first-order condition for an optimal t^*, which is therefore the only operative condition, is given by

$$\frac{d\mathcal{U}(\theta^*, t^*)}{dt} = -\frac{d\mathcal{V}(\theta^*, t^*)}{dt}. \tag{7}$$

In particular, the optimal t may be found by setting both sides of the last equation to zero. However, since neither α nor β levels are affected by changes in transfers, this equation continues to hold for all θ. All Pareto optima are associated with the same t^*, as depicted in figure 1.

Zero will be the optimal tax if

$$0 = \frac{d\mathcal{U}(\theta, 0)}{dt} \tag{8}$$

and

$$0 = \frac{d\mathcal{V}(\theta, 0)}{dt} \tag{9}$$

for some θ.

Let θ_1 and θ_2 be the solutions of (8) and (9). Then, $t^* = 0$ requires $\theta_1 = \theta_2$.

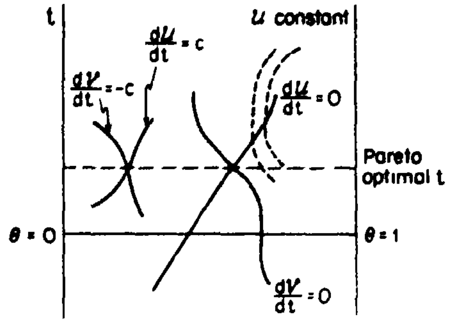

t
$\frac{d\mathscr{U}}{dt} = c$
$\frac{d\mathscr{V}}{dt} = -c$
$\mathscr{U}$ constant
$\frac{d\mathscr{U}}{dt} = 0$
Pareto optimal t
$\theta = 0$
$\theta = 1$
$\frac{d\mathscr{V}}{dt} = 0$

FIG 1

Differentiating the $\mathscr{U}$ and $\mathscr{V}$ totally with respect to t using the budget constraints obtained by substituting (3) and (4) into (1) and (2), and making use of the first-order conditions (5), one can derive

$$\bar{\theta}_1 = \frac{\alpha_1 - u_3(d\alpha_2/dt)}{\alpha_1 + \alpha_2} \tag{10}$$

$$\bar{\theta}_2 = \frac{\alpha_1 + v_3(d\alpha_1/dt)}{\alpha_1 + \alpha_2}, \tag{11}$$

where $d\alpha_i/dt$ is the total derivative of consumption with respect to t, including the change induced through an altered level of the externality, in equilibrium In the case of an externality that is harmful to both individuals, it should be noted that $\bar{\theta}_1 = \bar{\theta}_2$ will be satisfied if (exactly) one of the individuals increases his consumption of the externality-causing good in response to increases in the tax—that is, if $d\alpha_1/dt$ and $d\alpha_2/dt$ have opposite signs

One can regard the change in demand as being composed of two parts. First, there is a direct effect of price, holding the externality fixed, which amounts to a pure-own substitution effect (movement along a given demand curve) Since there are no income effects, this must be negative. Second, there is the effect of the altered level of the other individual's consumption at constant prices, the sign of which depends on the cross-derivative of utility with respect to these arguments. If a higher level of the externality induces a greatly increased level of demand, this may outweigh the first effect and produce a positive total derivative with respect to t.[1] The example of this section has this property, as will be seen below.

Differentiating the first-order conditions (5) with respect to the α's,

---

[1] This is related to the perverse behavior of aggregate consumption of the externality generating good between equilibria and Pareto optima, as discussed in Diamond and Mirrlees (1973).

β's, and t, holding p and q fixed with $z = 0$, and solving for $d\alpha_1/dt$ and $d\alpha_2/dt$, we have

$$\frac{d\alpha_1}{dt} = \frac{u_{22}\check{V} + v_{22}\check{U}}{\hat{U}\hat{V} - \check{U}\check{V}}, \qquad \frac{d\alpha_2}{dt} = \frac{u_{22}\check{V} + v_{22}\check{U}}{\hat{U}\hat{V} - \check{U}\check{V}}, \tag{12}$$

where

$$\hat{U} = u_{11}u_{22} - u_{12}^2,$$
$$\check{U} = u_{12}u_{23} - u_{13}u_{22},$$
$$\hat{V} = v_{11}v_{22} - v_{12}^2,$$
$$\check{V} = v_{12}v_{23} - v_{13}v_{22}.$$

Note that $\hat{U}$ and $\hat{V}$ are positive due to the second-order conditions. The terms $\check{U}$ and $\check{V}$ are related to the effects of the externality on the two goods in question

We have, therefore, that $\bar{\theta}_1 = \bar{\theta}_2$ if

$$u_3(u_{22}\check{V} + v_{22}\check{U}) + v_3(u_{22}\check{V} + v_{22}\check{U}) = 0 \tag{13}$$

Similarly, $\mathscr{U}'(\theta, z)$ and $\mathscr{V}'(\theta, z)$ are the utilities attained when only indirect taxation is allowed, and θ_1' and θ_2' are defined by

$$0 = \frac{d\mathscr{U}'(\theta_1', 0)}{dz} = \frac{d\mathscr{V}'(\theta_2', 0)}{dz} \tag{14}$$

The optimal z will be different from zero whenever $\theta_1' \neq \theta_2'$. Following the analysis above,

$$\frac{d\alpha_1}{dz} = \frac{u_{12}\check{V} + v_{12}\check{U}}{-\hat{U}\hat{V} + \check{U}\check{V}}, \qquad \frac{d\alpha_2}{dz} = \frac{u_{12}\check{V} + v_{12}\check{U}}{-\hat{U}\hat{V} + \check{U}\check{V}} \tag{15}$$

Condition (14) is therefore

$$u_3(u_{12}\check{V} + v_{12}\check{U}) + v_3(u_{12}\check{V} + v_{12}\check{U}) \neq 0, \tag{16}$$

where all derivatives are evaluated at the consumption levels of the untaxed equilibrium. Rewriting (13) and (16) we have

$$\check{V} = \frac{-v_{22}}{u_{22}}\check{U} - \frac{v_3}{u_3}\check{V} - \frac{v_3v_{22}}{u_3u_{22}}\check{U} \tag{17}$$

$$\check{V} \neq \frac{-v_{12}}{u_{12}}\check{U} - \frac{v_3}{u_3}\check{V} - \frac{v_3v_{12}}{u_3u_{12}}\check{U}. \tag{18}$$

To show that (17) and (18) can be consistent, let U and V have the forms

$$U = f(\alpha_1, \beta_1) + g(\alpha_2), \qquad V = h(\alpha_2, \alpha_1) + j(\beta_2). \tag{19}$$

Thus,

$$\tilde{U} = 0,$$
$$\tilde{V} = -v_{13}v_{22},$$
$$\tilde{V} = v_{11}v_{22}$$

Hence, conditions (17) and (18) are

$$\left(\frac{v_3}{u_3} v_{11} - v_{13}\right) = \frac{-\tilde{U}}{u_{22}}, \tag{20}$$

$$v_{22} \neq 0, \tag{21}$$

and

$$\tilde{U} \neq 0 \tag{22}$$

Since v_3, u_3 and v_{11} are all negative, we can show that these relations may be satisfied when evaluated at points that would be chosen in a competitive economy without taxes This is done by selecting a value for u_3 to satisfy (20), having fixed all of the other functions Because of the separable functional form of U, this can be done without disturbing either of the relationships (21) or (22) In order to be sure not to violate the nonpositivity of u_3 in doing so, we require

$$v_{13} < \frac{\tilde{U}}{u_{22}} \tag{23}$$

It is therefore necessary to show only that the functions f and h can be chosen so as to satisfy (23) Let us take

$$f(\alpha_1, \beta_1) = \alpha_1^{a_1}\beta_1^{b_1},\ a_1 + b_1 < 1;\ a_1, b_1 > 0, \tag{24}$$

$$h(\alpha_2, \alpha_1) = \alpha_2^{a_2}\alpha_1^{-c},\ 1 > a_2 > 0,\ c > 0 \tag{25}$$

One can show that

$$v_{13} = -cp\left(\frac{p}{a_1}\right)^{(2c+1)(1-b_1)/(1-a_1-b_1)}\left(\frac{q}{b_1}\right)^{(2c+1)b_1/(1-a_1-b_1)} \tag{26}$$

If $p > a_1$ and $q > b_1$, then v_{13} becomes arbitrarily large and negative as c becomes large Further, the equilibrium values of α_1 and β_1 are independent of c; hence $\tilde{U}/u_{22}$ is fixed Therefore we can choose c sufficiently large and will obtain the desired example

Two further remarks need to be made in this context. We have not shown that the only solutions to the first-order conditions are of the form

$t^* = 0, z^* \neq 0$. However, we have placed no restrictions on $u_3 = g'(\alpha_2)$ except at the point given by the choice of α_2 with zero tax rates on both commodities One can easily see that by choosing $g'(\alpha_2)$ appropriately at other values, it can be arranged that (19) is not satisfied at any nonzero t. Hence the local character of this example can be extended globally to imply that no direct commodity taxation can improve the laissez-faire equilibrium, although indirect methods are effective.

Second, one might want to know the sign of the tax (subsidy) in the market for β which is superior to any direct taxation on α Suppose the externality is harmful so that $u_3, v_3 < 0$ and that α and β are complements in the ordinary sense of a negative cross-elasticity of demand. Here we would expect a positive tax level to be the optimal indirect action, tending to reduce the level of the externality

A positive z will be optimal if, at $z = 0$, we have $(\partial \mathscr{U}'/\partial z) > -(\partial \mathscr{V}'/\partial z)$ where $\mathscr{U}'$ and $\mathscr{V}'$ are as described above.

This is equivalent to $\theta_1' < \theta_2'$. Or, following an analogous derivation to equations (10) and (11), this is,

$$-u_3 \frac{d\alpha_2}{dz} < v_3 \frac{d\alpha_1}{dz} \tag{27}$$

Substituting (15) in the above we have

$$u_3 \frac{u_{12}\breve{V} + v_{12}\breve{U}}{\breve{U}\breve{V} - \breve{U}\breve{V}} + v_3 \frac{u_{12}\breve{V} + v_{12}\breve{U}}{\breve{U}\breve{V} - \breve{U}\breve{V}} < 0 \tag{28}$$

Since $\breve{U} = 0$ in the above example, and $\breve{U}\breve{V} > 0$ by the second-order conditions, we have

$$u_3(u_{12}\breve{V} + v_{12}\breve{U}) + v_3 u_{12}\breve{V} < 0 \tag{29}$$

is the condition for z^* to be positive

Using (16), since $t^* = 0$ in the example, this becomes

$$\breve{U}\left(\frac{-v_{22}u_{12}u_3}{u_{22}} + u_3 v_{12}\right) < 0 \tag{30}$$

Since v_{12} is zero, this is surely violated in the example, and the optimal indirect tax has the perverse sign—positive for substitutes, negative for complements. However, more complex examples in which $v_{12} \neq 0$ may provide instances in which the sign of z^* has the expected behavior and indirect taxation is still superior This remains as an open question.

Positive Results

Having seen that indirect corrective taxation is superior under certain conditions, we now turn our attention to the classification of conditions under which such aberrations are impossible. The general problem of the superiority of direct taxation will not be treated here. Our more modest results concern instances in which a zero tax on the related commodity is best, the optimum being to operate solely on the market of the externality-causing good even though more general policies are available. We will show that this is the case if individuals are identical, or if the externality depends on the aggregate consumption of the good in question, each individual neglecting his influence on this aggregate. Again, certain remarks concerning nonconvexities must be made, but we will demonstrate some instances in which our conclusions hold without qualification

Let

$$\begin{aligned} U &= U(\theta, t, z), \\ V &= V(\theta, t, z), \end{aligned} \tag{31}$$

be the utility levels attained with taxes t and z and redistribution parameter θ By the linearity of utility in income, $W = U + V$ is independent of θ Necessary conditions for an optimal pair of taxes (t^*, z^*) are

$$\begin{aligned} 0 &= \frac{dW}{dt} = (v_3 + t)\frac{d\alpha_1}{dt} + (u_3 + t)\frac{d\alpha_2}{dt} + z\left(\frac{d\beta_1}{dt} + \frac{d\beta_2}{dt}\right), \\ 0 &= \frac{dW}{dz} = (v_3 + t)\frac{d\alpha_1}{dz} + (u_3 + t)\frac{d\alpha_2}{dz} + z\left(\frac{d\beta_1}{dz} + \frac{d\beta_2}{dz}\right), \end{aligned} \tag{32}$$

all derivatives being evaluated at the points chosen with taxes (t^*, z^*). One can solve these conditions for t and z at the optimum This yields

$$z^* = \frac{[(d\alpha_1/dz)\ (d\alpha_2/dt) - (d\alpha_1/dt)\ (d\alpha_2/dz)](v_3 - u_3)}{[(d\alpha_1/dt) + (d\alpha_2/dt)]\ [(d\beta_1/dt) + (d\beta_2/dt)] - [(d\alpha_1/dz) + (d\alpha_2/dz)]\ [(d\beta_1/dz) + (d\beta_2/dz)]}. \tag{33}$$

Substituting (12) and (15) in (33), one finds that

$$z^* = 0, \qquad \text{if } v_3 = u_3 \quad \text{or} \quad v_{12}u_{22} = u_{12}v_{22} \tag{34}$$

Two special instances in which these conditions hold are when individual's preferences are identical, or when they are separable in β In general, however, the optimal policy will involve taxes/subsidies in indirect markets Even in the case of identical utilities the foregoing conclusion depends on the fact that individuals will have identical consumptions in equilibrium

Fixing prices and taxes, we can think of each individual's choice of α as it depends on the level of the externality he faces This gives rise to two identical reaction functions, the intersections of which are equilibrium consumptions of α for this model. We wish to establish that this equilibrium is unique and occurs at a symmetric point. That is, we want to establish conditions for these reaction curves to be as in case (1) rather than case (2), in figure 2

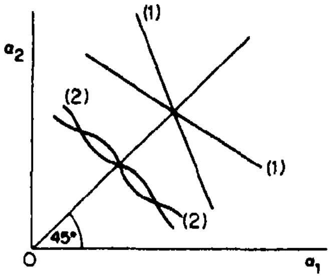


FIG 2

A condition for this uniqueness to hold is that the slope of the reaction curve be between 0 and -1 everywhere This will hold if utility can be written in the form

$$u[f(\alpha_i, \alpha_1 + \alpha_2), \beta_i], \quad i = 1, 2 \tag{35}$$

The interpretation of such a functional form is that the externality takes the form of aggregate consumption and affects the effective consumption of services α, but not β In this context it is natural to assume that simultaneous equal increases in the consumption of α by both individuals does not increase the effective level for both of them Such an upward variation in α_1 and α_2 ($d\alpha_1 = d\alpha_2$) changes f by $f_1 d\alpha_1 + f_2(d\alpha_1 + d\alpha_2)$ Therefore we require

$$f_1 + 2f_2 > 0 \tag{36}$$

The individual's first-order conditions are

$$\begin{aligned} u_1 f_1 &= p + t, \\ u_2 &= q + z. \end{aligned} \tag{37}$$

Differentiating totally, the slope of the individual's reaction curve is shown to be

$$\frac{d\alpha_i}{d\alpha_j} = \frac{f_2}{f_1 + f_2} \tag{38}$$

By (36) this is shown to be between 0 and -1 as desired

In the foregoing analysis we have assumed that the individual takes into account the full impact of his actions on his utility However, particularly in the case of a large number of consumers, it may be more natural to assume that consumption has direct benefits and also contributes to an externality-causing aggregate The individual may be ignoring his marginal contribution to this aggregate This form of externality has been termed "atmosphere" by Meade (1952), who found that direct taxation on the components of this aggregate could lead to a Pareto optimum In the context of the above analysis

$$\begin{aligned} U &= \phi(\alpha_1, \beta_1, \alpha_1 + \alpha_2) + X_1 + T_1, \\ V &= \psi(\alpha_2, \beta_2, \alpha_1 + \alpha_2) + X_2 + T_2. \end{aligned} \tag{39}$$

The assumption that individuals disregard their contribution to the externality is reflected in the equilibrium conditions

$$\begin{aligned} \phi_1 &= p + t, & \psi_1 &= p + t, \\ \phi_2 &= q + z, & \psi_2 &= q + z \end{aligned} \tag{40}$$

One should be careful in comparing (40) with the rather similar-looking conditions (5) The utility functions (1) and (12) are different from (39) in their behavioral implications because there the individual's own consumption of α affects his utility in a way that he takes account of explicitly In (39) and (40) he optimizes with respect to his choice of α by setting the derivative of utility with respect to its first argument equal to price but does not subtract the effect of α through the third argument. If the same behavioral hypothesis in (39) were taken, we would have had $\phi_1 + \phi_3 = \psi_1 + \psi_3 = p + t$, instead

Letting $W = U + V$ we find that

$$\begin{aligned} dW = {} & (\phi_1 + \phi_3)d\alpha_1 + \phi_2 d\beta_1 + \phi_3 d\alpha_2 + (\psi_1 + \psi_3)d\alpha_2 \\ & + \psi_2 d\beta_2 + \psi_3 d\alpha_1 - p(d\alpha_1 + d\alpha_2) - q(d\beta_1 + d\beta_2) \end{aligned} \tag{41}$$

Using the first-order condition (40),

$$dW = (t + \phi_3 + \psi_3)d\alpha_1 + (t + \phi_3 + \psi_3)d\alpha_2 + z(d\beta_1 + d\beta_2) \tag{42}$$

Thus $t^* = -\phi_3 - \psi_3$, $z^* = 0$ solves the optimization problem even though both direct and indirect tax policies are available

Following Diamond (1973), a more general form of atmosphere-type externalities is given by the utility functions

$$\begin{aligned} U &= \phi[\alpha_1, \beta_1, \gamma(\alpha_1, \alpha_2)] + X_1 + T_1, \\ V &= \psi[\alpha_2, \beta_2, \gamma(\alpha_1, \alpha_2)] + X_2 + T_2, \end{aligned} \tag{43}$$

for some function γ. As above, individuals are assumed not to perceive their influence on the level of γ, so that first-order conditions for their maximization problems are given by (40).

Differentiating W with respect to t and z, a necessary condition for the optimal z^* to be zero is

$$(\gamma_1 - \gamma_2)\left(\frac{d\alpha_1}{dz}\frac{d\alpha_2}{dt} - \frac{d\alpha_1}{dt}\frac{d\alpha_2}{dz}\right) = 0 \tag{44}$$

In the additive case $\gamma_1 = \gamma_2$, and (44) is always satisfied. This will also hold in the case of identical individuals in the absence of additivity (see above).

It will also hold, even if individuals are not identical, provided their relative changes in the consumption of the externality-causing commodity with respect to the two taxes are equal, that is

$$\frac{(d\alpha_1/dt)}{(d\alpha_1/dz)} = \frac{(d\alpha_2/dt)}{(d\alpha_2/dz)}.$$

More generally, if (44) holds at a point where

$$t^* = \frac{(\phi_3 + \psi_3)[(d\alpha_1/dt)\gamma_1 + (d\alpha_2/dt)\gamma_2]}{[(d\alpha_1/dt) + (d\alpha_2/dt)]}, \tag{45}$$

and $z^* = 0$, then this is an optimum. Notice that (44) and (45) imply that $z^* = 0$ only when

$$\frac{\gamma_1(d\alpha_1/dt) + \gamma_2(d\alpha_2/dt)}{(d\alpha_1/dt) + (d\alpha_2/dt)} = \frac{\gamma_1(d\alpha_1/dz) + \gamma_2(d\alpha_2/dz)}{(d\alpha_1/dz) + (d\alpha_2/dz)},$$

which means that the sum of marginal externalities, weighted by quantity response to direct price change (t) is equal to the sum of marginal externalities weighted by quantity response to indirect price change (z).[2]

References

Baumol, W. J. "On Taxation and the Control of Externalities." *A.E.R.* 62 (June 1972): 307–22.

Buchanan, J. M. "External Diseconomies, Corrective Taxes, and Market Structure." *A.E.R.* 59 (March 1969): 174–77.

Davis, O. A., and Whinston, A. "Externalities, Welfare, and the Theory of Games." *J.P.E.* 70 (June 1962): 241–62.

Diamond, P. A. "Consumption Externalities and Imperfect Corrective Pricing." *Bell J. Econ. and Management Sci.* 4 (Autumn 1973): 526–38.

Diamond, P. A., and Mirrlees, J. A. "Aggregate Production with Consumption Externalities." *Q.J.E.* 87 (February 1973): 1–24.

Johnson, H. G. *Aspects of the Theory of Tariffs.* Cambridge, Mass.: Harvard Univ. Press, 1972.

Meade, J. E. "External Economies and Diseconomies in a Competitive Situation." *Econ. J.* 62 (March 1952): 59–67.

Pigou, A. C. *The Economics of Welfare.* 4th ed. London: Macmillan, 1932.

Plott, C. R. "Externalities and Corrective Taxes." *Economica* 33 (February 1966): 84–87.

---

[2] We owe this observation to Peter Diamond.

Technology, Growth, and History

Jeffrey G. Williamson

University of Wisconsin

I. Theory and History

There are some economic problems for which historical analysis is especially well suited, indeed essential. Economic historians who fail to be motivated by this fact are doomed to have only the most limited and transitory impact on economic thought. Paul David is one of our best pathfinders in this entwined thicket of theory and history. The appearance of his book *Technical Choice, Innovation and Economic Growth* (1975) should help speed the diffusion of some highly original ideas on technological change, distribution, and economic growth. David is an economic historian—or better, an economist and a historian—who hammers on the toughest theoretical nuts and who recognizes the true value of history in shedding light on economic dynamics.

The book is separated into two parts. A very long chapter 1 deals with technical choice, innovation, and technological change at very high levels of aggregation. Here David writes macrohistory in the grand style and emerges grasping an aggregate model of growth with the bias and the rate of technological change endogenous. If it held water, it would be an achievement of some note. I have doubts, and they are presented at the end of this review in Section IV. They appear last in the review since they reflect on Professor David's thoughts of most recent vintage. The majority of the book's pages, however, are devoted to "micro-historical" essays published between 1966 and 1973, two of which have already become classics. Chapters 2 and 3 deal with learning by doing as an endogenous source of efficiency improvement.[1] Section II will

---

Although I accept full responsibility for its contents, this review has been improved by the critical comments of Don Devoretz, Peter Lindert, Samuel Morley, Alan Olmstead, Nathan Rosenberg, George Stigler, and Vernon Ruttan.

[1] Chapter 2 was originally published as David (1970), although an "Addendum" (pp. 169–72) on labor-quality estimates has been added.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved.

attempt a critical evaluation of this material. Chapters 4 and 5 deal with the diffusion of new technology,[2] and these are the subject of the commentary in Section III.

II. Microhistory: Learning by Doing and Machine Quality

When "Learning by Doing and Tariff Protection" first appeared in 1970, it was considered by many one of the finest examples of empirical research on micro aspects of technical change. Reading it again as chapter 2 has failed to change that initial evaluation, at least for me. In spite of the enormous influence which the learning hypothesis has had since Arrow (1962) made his formal restatement, few hard tests of the thesis exist. Certainly no economic historian had yet evaluated the quantitative importance of the thesis in accounting for productivity performance.

David's contribution is extraordinarily ambitious. First, it contains a test of various forms of the learning-by-doing hypothesis using the antebellum cotton textile industry to generate his sample. Two versions of accumulated experience (Fellner 1969, p. 124) are utilized, and the econometric results allow David to discriminate in favor of one over another: cumulated aggregate output, David asserts, fails when compared with the temporal duration of production experience. Second, the results of that test are applied to the conventional infant-industry argument. What makes David's empirical results so interesting is that protectionists have long viewed American antebellum cotton textiles as the classic example in support of the infant-industry argument. The econometric results ". . . imply that from a technical standpoint the blanket subsidies provided for all domestic entrants into cotton textile production were—from the Tariff of 1824 onward—a largely redundant set of measures, essentially the same benefits, such as they were, might have been secured under a Free Trade regime by policies which selectively subsidized the accumulation of transferable knowledge deriving from the commercial operation of pilot-learning enterprises" (p. 168). Third and most important, David then uses the estimated model to decompose the sources of productivity growth in this key industry during America's most critical decades of industrialization.

How important were experience-linked improvements in efficiency to New England's leading firms "in the years (1834–60) following the era of major technical and organizational innovations" in the cotton textile industry (p. 160)? We are told that learning accounted for eight-tenths of the annual total factor productivity growth during the 1830s. It was nearly three times as important as the rise of capital intensity (p. 164).

---

[2] These chapters are both based on papers in edited volumes. Chapter 4, the famous reaper paper, was first published in David (1966), while chapter 5 appeared in David (1971).

Furthermore, the retardation in subsequent productivity growth up to the late 1850s is attributable entirely to diminished learning effects (p. 164). Were these findings supported by further research, they would be enormously valuable to our theories of growth and technological progress.

Given the major role played by machine quality elsewhere in the theoretical and historical literature, David's result delivers quite a jolt. Indeed, his approach appears to clash even with a key portion of the learning-by-doing literature itself. Kenneth Arrow (1962) thought that gross investment was the relevant index of experience, and that "production experience in 'machine-building,' rather than in 'machine-using' " (p. 105) was what mattered. David does not test Arrow's thesis, but takes instead an antithetical position. Since the issue is central to the remainder of the book, we may be forgiven amplification and detail. While David's basic model attempts to adjust for changing labor quality, no such adjustment is made for capital. Indeed, he uses spindles in operation as a proxy for capital services. To equate an 1834 "spindle" to an 1860 "spindle" requires the adoption of a very heroic view of technical change. David assumes that the major innovations were introduced by 1834, and that new capital goods for expansion or replacement were basically the same thereafter. To the extent that historical reality was otherwise, the estimated learning coefficient is overstated. It follows that David's heroic, comparative static view of machine and organizational improvement almost inevitably yields the high learning effects estimated.

It seems evident that future research demands more careful attention to the distinction between learning from machine using and learning from machine building. No doubt it was precisely this concern that led Professor David to write chapter 3.[3] The Lawrence No. 2 Mill was built at Lowell in 1834 by the Lawrence Cotton Textile Company and is part of the sample of New England firms analyzed in chapter 2. Between installation and 1857, "there appears to have been no significant machinery replacement" and "no new investment in . . . this mill's capacity" (p. 176). While many improvements were made on the mill floor, Lawrence No. 2 seems to conform to the classic short-run Horndal case—productivity improvement in fixed facilities. David's analysis of Mill No. 2 assumes great importance since machine quality changes are minimized (though not eliminated) in this fixed-facilities case. David estimates $\ln(H/Y)_t = \alpha - \lambda \ln(T)_t$, where (H/Y) measures labor requirements per unit of output and T is accumulated experience (years) since 1834. This regression uses four annual observations—1839, 1846, 1851, and 1856—since these years are considered (the only?) full-capacity observations. The estimated short-run learning parameter is found to be smaller than the long-run parameter estimated in chapter 2, 0.23 versus 0.33 (p. 185). In other words, vintage of textile plant does count, thank the Lord, and changing machine quality does matter.

---

[3] Chapter 3 first appeared under the same title in David (1973).

Although they have some obvious flaws, these two chapters suggest a promising methodology to disentangle the relative importance of learning by using machines and learning by building machines. Both are endogenous historical forces, but they have sufficiently different implications to warrant much more intensive research.

III. Microhistory: Adoption, Diffusion, and Machine Quality

Obed Hussey and Cyrus McCormick demonstrated their reaping machines in 1833 and 1834 The demonstrations were successful, as our elementary history textbooks tell us in heroic terms While the reaper did eventually revolutionize the life of grain farmers, at least at harvest time, the adoption and diffusion were very slow during the 1830s and 1840s. These two decades were followed by an extraordinary burst of acceptance, and reaper production soared in the 1850s. Why did it take two decades for farmers to start replacing wholesale their cradles and scythes for reapers and mowers? The diffusion of harvesting machinery in America was an important economic event deserving the attention which historians have lavished on it over the years Yet none of his predecessors attacked the problem with the analytical vigor contained in David's 1966 piece, "The Mechanization of Reaping in the Ante-Bellum Midwest," reprinted here as chapter 4 The contribution was especially attractive since it offered a model capable of explaining the classic S-shaped diffusion pattern Furthermore, the model appeared to be testable and could be used to decompose the proximate forces mainly responsible for the burst of acceptance in the 1850s Was it the boom in grain prices on the product demand side? Improved machine quality and diminished price? Increased labor scarcity? Credit conditions? Size of farm? It is important to sort out these forces if we are to enrich our knowledge of technology diffusion David supplies the following conventional neoclassical answer· "The story of the adoption of the mechanical reaper in the years immediately before the Civil War should thus be told in terms of the effects of both an expansion of grain acreage sown on individual farms and the downward movement of the threshold size as a result of the rising relative cost of harvest labor. But, of the two types of adjustment taking place during the 1850's, the former must properly be accorded lesser emphasis" (p 216). This has a nice neoclassical ring to it. And where did the scarce labor pressure come from? It came from a midwest supply response induced by a boom in grain prices.

Whatever happened to our heroes Hussey and McCormick? The reaper is demonstrated in 1833 and 1834; from then on the supply side

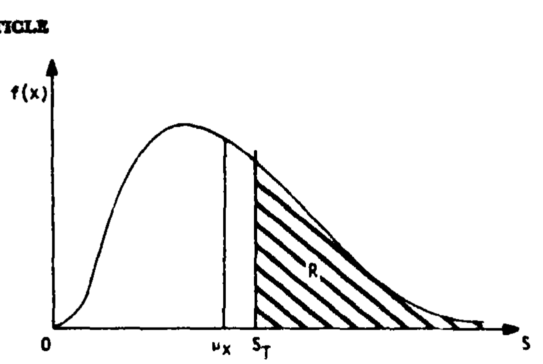


FIG 1

of the reaper market plays a passive role, awaiting the right demand forces. And when demand spreads nationwide, scale economies insure a supply of even cheaper "standard" (1833–34 model?) machines: "For the country as a whole, as well as for the Midwest, this afforded economies of scale in the production . of reaping machines. It thereby [ensured that] the long-run aggregate supply schedule for harvesting machinery was more elastic than the farm labor supply" (p 219). What is the model by which the supply side of the reaper market is limited to this walk-on part, while demand hogs the stage?

The key concept in David's model is the "threshold" farm size. A farmer will purchase a reaper if its annual user cost is less than the annual savings in wage costs attributable to the reaper. The threshold farm size, S_T, is defined by $c = S_T w L_S$, where L_S are man-days of labor (per acre harvested) saved at some daily wage rate, w Credit conditions, depreciation rates, maintenance costs, and purchase price all enter into the user cost of the reaper, c. Its quality, durability, and suitability to the topology determine the extent of labor saved per acre In any case, farmers endowed with farm acreage greater than S_T purchase the machine, while smaller farms do not If the frequency distribution, $f(x)$, of farms by acreage is known (fig 1), then by calculating S_T we can predict the proportion of farms which should have adopted the reaper Only the average, μ_x, farm size was available to David, not the distribution. He assumed that changes in the $(S_T - \mu_x)$ "gap" were excellent proxies for the behavior of R, and as empirical support for the model cites the fact that during the early 1850s "the threshold point dropped from over 46 to roughly 35 acres" (p. 215), while "a typical farm in [Illinois's] leading small-grain regions increased from 25 to 30 [acres]" (p. 216). The final step in David's accounting is to allocate the sources of the declining $(S_T - \mu_x)$ gap into (1) a rising μ_x and (2) a falling S_T and its causes.

The assumptions underlying David's model are very strong. Indeed, it appears that "machine-supply" forces have been relegated to a walk-on part by assumption rather than by historical evidence. In a very important paper on reaper adoption in Ontario after 1850, Pomfret (1974) shows that David's conclusions are reversed. An analysis limited solely to (S_T − μ_x) would have suggested that changes in factor prices dominated changes in average farm size in accounting for the Ontario adoption rate (Pomfret 1974, p. 11). Yet, when the known distribution of farm sizes is introduced, Pomfret finds that "... David's methodology would have produced a qualitatively wrong answer to the question of which of these two explanatory variables was the more important.. " (p. 12). Furthermore, Pomfret's model is quite capable of generating S-shaped diffusion paths too! Indeed, "Analysis utilizing the size distribution of innovating units sheds new light on the diffusion process and [since] firms in many industries are lognormally distributed, this method of analysis has potentially wide applicability."[4]

The reader will note that Pomfret uses the term "firm" rather than "plant" in the above quote. If contiguous working farms ("plants") were willing to share the reaper or mower (behave like an integrated "firm"), would the threshold concept lose its empirical relevance? How divisible was the reaper? Is there evidence of informal, cooperative sharing including family or neighbor joint ownership? David devotes surprisingly little attention to this fundamental issue, except to assert in a footnote that sharing was not common practice (p 208). Of course "common" practice would not have been necessary to insure its importance Importance would have been insured if the practice was typical among farms bunched below S_T in a given region. So crucial an assumption requires far more serious attention A recent paper by Alan Olmstead (1975) supplies just such detail It documents sharing frequency completely at odds with David's assertion It appears that the issue of sharing, contracting, and renting should have been resolved prior to David's application of his diffusion paradigm to nineteenth-century American farm mechanization The data are there, Olmstead and others show that archival research can readily resolve these crucial historical issues. It is a pity that David failed to do so some 10 years ago.

Chapter 5 extends the diffusion model to Victorian Britain. The model is less novel but more relevant in explaining British resistance to harvest mechanization from the 1840s to World War I Capital-intensive cereal production adopted during England's previous "agricultural revolution" produced hedges, fences, and field layouts which were simply incompatible with nineteenth-century labor-saving machinery. Simple enough, but in David's hands the disadvantages of early industrial leadership become more concrete with this attention to the technical interrelatedness between modern machines and long-lived physical plant.

---

[4] Pomfret (1974, p. 16). I should add that the concept is not completely novel, even in the literature on technology in agriculture.

A final point deserves some stress. These two chapters require an extravagant expression of faith in comparative statics, a religious commitment even stronger than that underlying his learning-by-doing analysis. Professor David assumes that a reaper is a reaper is a reaper—that there were no "significant" changes after 1833 in what David denotes the "standard" antebellum machine (Olmstead 1975, p. 329). While this makes the diffusion narrative more readable, its accuracy may be doubted. By so assuming, David ignores the equally important forces of quality improvement and the declining relative price of quality-adjusted machines Judged by Rosenberg's (1972) overview, David's assumption is almost always made by economists confronting the diffusion issue, but there is absolutely no evidence to support it On the contrary, abundant qualitative documentation of the "basic reaper's" improvement from a heavy, primitive device, hard to transport and prone to very frequent breakdown, to a more maneuverable, lighter machine requiring less maintenance and fewer operators, can be found in the classic accounts by Rogin, Ardrey, and Hutchinson A striking quantitative example is supplied by Pomfret (1974, p. 4), whose careful attention to quality changes in Canadian farm machinery offers quite a different evaluation of the forces influencing nineteenth-century North American farm mechanization The improvement in the quality of machines gets top billing, increased farm size ranks second, while "labor scarcity" drops to last place Is there any reason to believe that a recalculation of the American midwestern experience between the late 1840s and early 1850s would conflict with Pomfret's Ontario story for the 1850s and 1860s? Surely there were machine improvements in the early American hand rake (Olmstead 1975, pp 344–51), and the quantitative importance of those "secondary inventions" is an empirical issue which no amount of theorizing will resolve

IV. Macrohistory: Labor Scarcity and Technological Progress

Chapter 1 contains totally new material and presents a breathtaking challenge to the reader It is long, difficult, and demanding Before the reader can begin to comprehend, appreciate, and critically evaluate these 90 pages, he must grasp the microanalytic ideas of the later chapters, he must have an excellent grounding in the recent theoretical literature on growth and technical change, he must have a working knowledge of the nineteenth-century historical facts, and he must have read the recent historical literature—cliometric or otherwise Paul David is a Renaissance economist drawing on more than a decade's reflection on the dynamics of the macrohistorical process I find myself in frequent disagreement with the chapter, but never bored.

Three issues have always been central to empirical and theoretical studies of growth, development, and distribution. (1) The rate of technological progress· Why is the rate of productivity growth higher for some economies than others, for some sectors than others, for some historical eras than others? (2) The factor-saving bias of technological progress· Why is the rate of labor saving more pronounced in some nations than others, in some sectors than others, in some historical eras than others? (3) The choice of technique. What are the techniques available at any point in time and what determines the ultimate selection? Conventional economists are accustomed to partitioning these three issues into distinctly separate optimization problems, each capable of independent analysis. David rejects this approach and searches for a framework of historical dynamics which can deal simultaneously with all three

What are the historical facts? Interest in the "stylized facts" goes back at least as far as the 1840s when European visitors viewed American technology with amazement The British observers were the most lucid and prolific They directed their comments to the three central issues (p 21) In modern vernacular, America adopted more capital intensive technologies, had a more dramatic Hicksian labor-saving bias, and registered more rapid rates of total factor productivity growth Casual contemporary empiricism of this sort was sufficient to build Habakkuk's (1962) classic *American and British Technology in the Nineteenth Century* Oddly enough, in spite of the justifiable attention which Habakkuk's book has attracted in the years since, we do not have much more in the way of hard nineteenth-century facts, especially regarding the bias (Asher 1972) However, the twentieth-century macroeconometric evidence seems to be consistent with the labor-saving characterization

Scarce labor would surely explain the more capital intensive technologies in America, but what about labor saving? Fellner's (1961) analytical solution was to drop the perfectly competitive assumption, to rely on differentially imperfect factor markets, and thus to stress entrepreneurial anticipation of future changes in the relative marginal costs of their inputs Given Habakkuk's (1962) characterization of U S labor supply as being relatively inelastic, the countervailing labor-saving bias of new technologies was rationalized David is unhappy with this approach primarily because American natural resource abundance fails to play a convincing role (pp 33–39)

David's alternative hypothesis begins to emerge in pages 62–68, and a simplified version appears as figure 2 Clarity is achieved by restricting our attention to two inputs and two techniques. The β-ray is always more capital intensive than the α-ray Current engineering knowledge is described by APF^{0}, the "available process frontier." Thus, England with cheap labor and facing an input price line pp selects B_0 along the labor-intensive α-ray America with its dear labor ($p'p'$) selects A_0 on the capital-intensive β-ray. So far, this is nothing new. But note that the

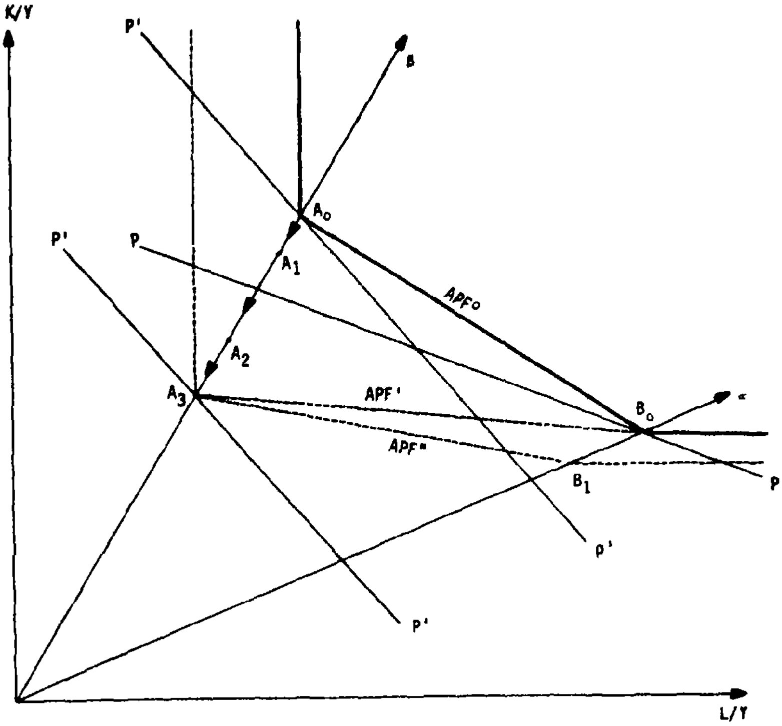


FIG 2

historical conditions which produce $p'p'$ and thus the technique choice at A_0 now start in motion a dynamic irreversibility in the global bias. In 18 pages (pp 68–86) of sophisticated appeal to random walks and compulsive sequences, David presents an analytical justification for expecting

> . . an incidental, myopic exploration of the β-ray But as this learning process proceeds, the power of price variations to halt the emerging labor-saving drift of the APF begins to diminish. Once the point of the β-ray's intersection with pp is reached, a mere restoration of the *status quo ante* in the factor markets would surely not draw even the most myopic producer back to the α-technique Eventually, even with progress down the β-ray occurring at a retarded rate—as we generally expect to happen on a learning curve—the APF would become approximately L-shaped . . . There is thus some theoretical basis for seeking the origins of the modern configuration of a society's technology in the accidents of its *remote* factor-price history. [Pp 66–67]

Figure 2 conforms to the stylized facts: (1) America is more capital intensive at A_3 than Britain at B_0; (2) the rate of labor saving is more dramatic in America as she moves from A_0 to A_3, while Britain remains stationary at B_0; (3) America's rate of total factor productivity growth is higher Why doesn't Britain explore her α-ray? Presumably because she has already completed her learning-by-doing sequence. This is an extreme view, of course, taken only for emphasis; the same point can be made by allowing Britain to myopically search her α-ray and to move from B_0 to B_1 Why does America move more rapidly along β than Britain does along α? America exhibits higher total factor productivity growth because her learning-by-doing rate is more dramatic According to "Habakkuk's theorem," the scope for learning is greater simply because there is more to learn about a machine-intensive technology The "theorem" is critical to David's macrohistory, it has yet to be tested.

What is the accident of America's "remote factor-price history" which starts this unique and irreversible dynamic process? The answer lies with natural resource abundance, a conventional answer perhaps, but the reasons are unconventional " . relatively capital-intensive techniques were also relatively resource-using" (p 88). David's macrohistory is now complete " even if the same labor-capital price ratio (*pp*) had faced producers in Britain and America, the comparatively greater availability of natural resources would have suggested to some American producers the design, and to others the selection for use, of more capital-intensive methods Technique β would be chosen in preference to methods like α And from such choices immediately would flow the consequences for the global characteristics of the ensuing technological developments in each country, as has already been detailed" (p 90)

V. Reprise

Most of my difficulties with David's macrohistory stem from his curious disregard for commodities, industries, and sectors There is no industrialization in David's book The demise of the farm, the rise of the city, western settlement, unbalanced rates of output and total factor productivity growth—these elements are totally absent from David's account I think his narrative is weaker as a result

Years of macroeconometric work have established evidence of labor saving, at least in twentieth-century America. David cites much of this work but fails to note that Morishima and Saito (1968) attribute the vast majority, if not all, of (global) labor saving to a shifting output mix toward capital intensive (machines cum skills) commodities This raises a set of questions which are completely ignored in David's narrative·

What role does commodity demand play in determining output mix, and thus an economy's position on APF* (or some appropriate contract curve), and thus subsequent historical dynamics? Could it be that progress in income per capita, whatever its causes, generates its own (aggregate) labor-saving bias via Engel's law? Could it be that unbalanced rates of (neutral) total factor productivity growth, high in modern capital cum skill-intensive sectors and low in traditional unskilled labor-intensive sectors, insure a shift into cheapening modern products, thus generating an aggregate labor-saving bias? My own feeling is that this is the stuff of macrohistory, that general equilibrium analysis is essential for such problems, and that the partial equilibrium graphics of figure 2 are simply not up to the task.

Furthermore, David's account relies too heavily on learning-by-using machines There is no attention to the machine-building sector How can we write a macrohistory without explicit attention to the capital goods sector? The extraordinary plunge of American machine prices (Williamson 1975, pp. 2–5, 103–12) cannot be ignored in any such history This American success story needs explanation because, at least in my view, herein lies the key to America's choice of machine-intensive technologies and machine-intensive commodities I suspect the explanation will require greater attention to labor of different skill and scale economies in capital goods production (Brito and Williamson 1973)

In spite of such criticisms, the reader can plainly see that Paul David asks very difficult and extremely interesting questions indeed. *Technical Choice, Innovation and Economic Growth* is a stimulating book. One can only hope that more cliometricians will follow David's lead and increase their historical stake in gardens which promise yields equally relevant to mainstream economic thought

References

Arrow, K J "The Economic Implications of Learning by Doing." *Rev. Econ Studies* 29 (June 1962) 155–73

Asher, E. "Industrial Efficiency and Biased Technical Change in American and British Manufacturing The Case of Textiles in the Nineteenth Century." *J Econ Hist* 32 (June 1972) 431–42

Brito, D L., and Williamson, J. G. "Skilled Labor and Nineteenth Century Anglo-American Managerial Behavior " *Explorations Econ. Hist* 10 (Spring 1973): 235–52.

David, P. A. "The Mechanization of Reaping in the Ante-Bellum Midwest." In *Industrialization in Two Systems*, edited by H. Rosovsky New York· Wiley, 1966

———. "Learning by Doing and Tariff Protection A Reconsideration of the Case of the Ante-Bellum United States Cotton Textile Industry." *J. Econ. Hist.* 30 (September 1970): 521–601.

———. "The Landscape and the Machine: Technical Interrelatedness, Land Tenure and the Mechanization of the Victorian Corn Harvest." In *Essays on a Mature Economy: Britain after 1840,* edited by D. N. McCloskey. London: Methuen, 1971.

———. "The 'Horndal Effect' in Lowell, 1834–1856: A Short-Run Learning Curve for Integrated Cotton Textile Mills." *Explorations Econ. Hist.* 10 (Winter 1973): 131–50.

———. *Technical Choice, Innovation and Economic Growth: Essays on American and British Experience in the Nineteenth Century.* Cambridge: Cambridge Univ. Press, 1975.

Fellner, W. "Two Propositions in the Theory of Induced Innovations." *Econ. J* 71 (June 1961): 305–8.

———. "Specific Interpretations of Learning by Doing." *J. Econ. Theory* 1 (August 1969): 119–40.

Habakkuk, H. J. *American and British Technology in the Nineteenth Century.* Cambridge: Cambridge Univ Press, 1962

Morishima, M., and Saito, M. "An Economic Test of Hicks' Theory of Biased Induced Inventions." In *Value, Capital and Growth,* edited by J. Wolfe. Chicago: Aldine, 1968

Olmstead, A. L. "The Mechanization of Reaping and Mowing in American Agriculture, 1833–1870." *J Econ Hist.* 35 (June 1975): 327–52.

Pomfret, R. "The Introduction of the Mechanical Reaper in Ontario, 1850–70." Mimeographed. Simon Fraser Univ, Dept Econ, 1974.

Rosenberg, N "Factors Affecting the Diffusion of Technology." *Explorations Econ. Hist* 10 (Fall 1972): 3–34.

Williamson, J. G. *Late Nineteenth Century American Development: A General Equilibrium History.* Cambridge: Cambridge Univ Press, 1975

Communications

The Friedman-Savage Utility Function in Cross-cultural Perspective

Frederic L. Pryor

Swarthmore College

A. Introduction

In their classic article on the utility of wealth, Friedman and Savage (1948) tried to devise a utility function to explain a variety of apparently conflicting phenomena. For instance, why do people who normally do not play gambling games such as calling heads-or-tails participate in lotteries? And why do people who gamble also buy insurance?

The purpose of this note is to show that the ideas underlying the Friedman-Savage utility function can permit us to make accurate predictions about which precapitalist societies do and which do not engage in gambling. The supporting data are macroeconomic, and the tests are carried out using a simple multivariate least-squares regression analysis.

B. The Utility-of-Wealth Function

The Friedman-Savage utility function has a rather strange shape, which is shown in figure 1 below. The utility function is the curve U–U'. A person normally has wealth W_n and experiences a utility U_n. He now faces two different gambling situations, in both of which his expected wealth after the bet (i.e., the sum of the probability of winning multiplied by his wealth position if he won plus the probability of losing multiplied by his wealth position if he lost) is the same as his normal wealth situation.

Some of the costs of writing this essay were financed by the Swarthmore College Faculty Research Fund, for which I express my gratitude. Useful comments on an earlier draft were graciously supplied by Van Doorn Ooms, Steven Piker, and Bernard Saffran.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved.

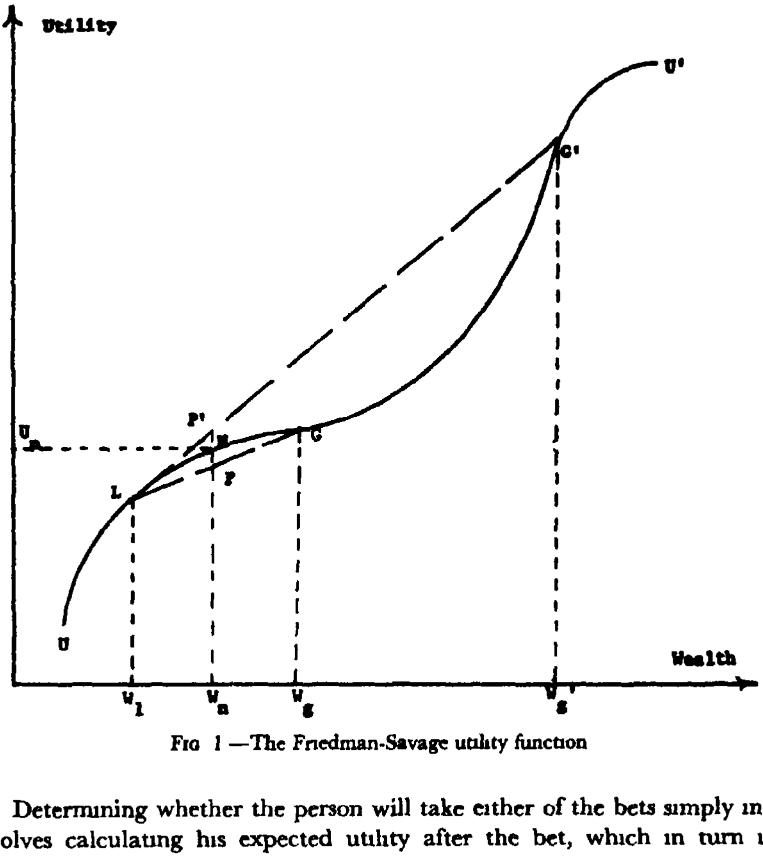


FIG 1 —The Friedman-Savage utility function

Determining whether the person will take either of the bets simply involves calculating his expected utility after the bet, which in turn is determined as the sum of the probability of winning multiplied by the utility of his wealth position if he won plus the probability of his losing multiplied by the utility of his wealth position if he lost. In the first bet, the person will have wealth position *L* if he loses and wealth position *G* if he wins The weighted utilities of these two wealth positions falls somewhere on the line *L–G* and, given the odds mentioned above, this is at point *P* (where the line indicating the expected wealth outcome intersects line *L–G*). Since point *P* has a utility less than U_n, the person will not take the bet. In the second bet, the person will have a wealth position *L* if he loses and a wealth position *G′* if he wins. The weighted utilities of these two wealth positions falls somewhere on the line *L–G′*, and, given the odds mentioned above, this is at point *P′*. Since point *P′* has a utility *more* than U_n, the person will take the bet.

Friedman and Savage tried to interpret the derived utility curve in terms of the entire society, an interpretation which was subjected to attack by Markowitz (1952). Nevertheless, the basic shape of this curve has been accepted for more than a quarter of a century.

Empirical verification of this function has rested, as far as I have been able to determine, exclusively on microeconomic data. My macroeconomic approach extends a number of microeconomic propositions, and I focus my analytic attention particularly on the wiggle in the curve and the points left of point G'. I do not deal with situations in which the curve begins to show a marked diminishing marginal utility again to the right of point G'.

C. The Test Situation and Alternative Hypotheses

What kind of determinants can be used to predict the presence of gambling in a worldwide sample of 60 primitive and peasant societies? I draw several predictions from the Friedman-Savage utility function and then discuss a series of alternative hypotheses found in the anthropology literature.

Before the Friedman-Savage utility function can be used as a basis for prediction, we must inquire briefly into the authors' underlying justifications of the rising marginal utility of wealth. Most of their discussion is in terms of rationalizing the behavior of individuals who both gamble and take out insurance. However, such behavior is based on certain conditions in society. They suggest that the two convex segments correspond to "qualitatively different socioeconomic levels, and the concave segment to the transition between the two levels. On this interpretation, increases in income that raise the relative position of the consumer unit in its own class but do not shift the unit out of its class yield diminishing marginal utility, while increases that shift it into a new class, that give it a new social and economic status, yield increasing marginal utility" (pp. 298–99). The greater the possibility of achieving such qualitatively different statuses, the greater the possibility of the presence of gambling. We can predict, therefore, that there should be a positive correlation between socioeconomic inequality and the presence of gambling in the various societies in the sample, a hypothesis validated below.[1]

Another implication can also be drawn. Individuals in societies where members feel themselves close to the subsistence level of income should have a very high disutility of losing since (if they have average incomes) losing might bring them close to starvation, therefore, the presence of gambling should be inversely correlated with this condition. Unfortunately, the measurement of income level raises some difficult problems.

---

[1] A hidden assumption in this argument is that social-class mobility is difficult and that gambling is one of the few easy channels of mobility. It should be added that this argument does *not* rest on any assumption about the utility of any given amount of wealth being the same for the individual societies or the utility functions being similar. Rather, I am assuming only that the utility-of-wealth functions have the same general shape (two segments with diminishing marginal utility of wealth, separated by a segment with an increasing marginal utility of wealth.)

Sahlins (1972) presented convincing evidence that highly primitive hunting and gathering societies are nowhere near this subsistence point, as shown by the fact that members of most of these societies spend very little time searching for food For instance, according to Lee (1968), the !Kung Bushmen of Botswana, who are among the most primitive peoples of the world, spend less than 20 percent of their work time looking for food Boserup (1965) has presented more controversial evidence, showing that the most primitive agricultural groups who employ slash-and-burn agricultural techniques spend much less time farming than do farmers using much more "advanced" techniques Since the food supply of these highly primitive societies can be increased by gathering natural products for a few more hours of the day, they should not feel close to the subsistence limit. Therefore, this subjective condition should not be correlated with the actual level of economic development, at least as conventionally measured by anthropologists such as Carneiro (1973) or Murdock and Provost (1973)

It does seem reasonable to suspect that the long-term feeling of closeness to subsistence occurs where the yearly food supply is quite uncertain. One such situation occurs in "hard" environments such as arctic, semi-arctic, or desert environments, where climatic shifts are considerable and the food quest is very difficult Another such situation occurs in that type of subsistence production in which the maintenance of a critical minimum capital stock is quite uncertain at all times for the various individuals in the society Of the five modes of food production (hunting, gathering, fishing, herding, and farming), it has been argued (e g , by Collins 1965) that such uncertainty is greatest in animal husbandry, where the herder must keep the size of the herd above a critical minimum and growing as fast or faster than the food needs of the herding family. In nomadic herding societies, where the entire food supply comes from the herds and a herd is in many different places during the course of the year, the dangers that the size of the herd may be reduced below the critical minimum by disease, theft, stampede, climatic occurrences, or accidents seem sufficiently great that gambling losses may have a very high disutility. Therefore, from considerations underlying the Freidman-Savage discussion, we can predict that the presence of gambling should be inversely related to hardness of the environment and also to reliance on nomadic herding; positive evidence for these predictions is presented below

In the anthropology literature, discussion about the determinants of gambling is much different and considerably less rigorous. Some of the studies are inductive correlation exercises (e.g , Roberts and Sutton-Smith 1966), while others attempt to trace the determinants of gambling to certain psychological variables such as values or attitudes induced by child-rearing practices. More interesting propositions come from various ethnological studies in which authors are trying to explain the presence of gambling in the societies they are investigating.

One common hypothesis from this latter literature is that gambling is related to the level of economic development More specifically, the higher the developmental level, the greater the individualism in the society and also the greater the breakdown of "strict tribal morality," two factors which might contribute to violations of injunctions against gambling. In addition, it is argued that economic development is related to the domestic use of money for commercial purposes and that such money serves to facilitate gambling We can test not only this relationship between economic development and the presence of gambling but also particular aspects of the argument For instance, we can test whether gambling is related to individualism by looking at proxies for individualism such as presence of nuclear (rather than extended) families and the absence of important lineage structures We can also test whether the presence of money influences the presence of gambling

Another kind of argument found in the anthropology literature rests on the possibility of cultural diffusion and comes in two versions. Version one is quite common gambling arises from the cultural degeneracy which allegedly occurs upon extensive contact with the West We can easily devise a measure for contact with the West, unfortunately, I expended much effort in coding various societies at periods in which such contact with the West was minimal, so that it is difficult to test this hypothesis satisfactorily with my data Although, in fact, the calculated regression coefficient for the contact variable is not statistically significant, a certain verification is found by looking at particular cases For instance, it is worth noting that at least three societies in my sample, namely the Truk Islanders of the Pacific, the Siane of New Guinea, and the Callinago Indians of the Caribbean, apparently did not aboriginally have gambling but adopted such practices after extensive Western contact.

A second version of the diffusion argument arises from an observation by the famous anthropologist Robert Lowie (1934, p 169), who observed that gambling is especially prevalent among the Indians of North America (north, that is, of Central America) He did not specify whether gambling was a trait brought by the original settlers of North America thousands of years ago and maintained by the various groups which split off to form separate societies or gambling was diffused more recently Nevertheless, I included a special location variable to take care of this contingency.

A final argument which friends of mine have suggested extends another intrasocietal argument to an intersocietal level in some societies most members are "risk lovers" and therefore practice gambling in order to fulfill this personal and cultural need Manifestations of "risk loving" are reliance on modes of subsistence in which there is a high daily variation in supply (e.g., hunting and fishing) or where the environment is "hard".

Thus the risk-loving hypothesis predicts that gambling will be positively correlated with the hardness of the environment, while the economic considerations discussed above predict exactly the opposite relation. It turns out that this risk-loving hypothesis provides very poor predictors of the presence of gambling.

D. The Sample

There are 60 primitive and peasant societies in the sample used for investigating the hypotheses discussed above, and they were drawn from all over the world. Murdock (1957) has divided the world up into 60 cultural areas, and 58 of these are represented in the sample. More than 85 percent of the 1,770 pairs of societies are more than 2,400 miles away from each other. Each society was coded for a particular year, usually in the second half of the nineteenth century, before much Western contact had occurred, however, this did not prove possible for some of the societies in the sample.

The data for coding the societies were drawn from more than 1,200 ethnographic sources, and this project represents a very small piece of a much larger comparative study of distribution in primitive and peasant societies (Pryor, forthcoming). My codings for the presence of gambling can be checked against the codings of Murdock and his associates for the presence of games of chance, and considerable agreement is obtained.[2] Difficulties in the coding of the independent variables was minimized in several different ways. First, only one coder (myself) made the various ratings, so that problems of intercoder reliability did not arise. Second, the data on the independent variables were collected for a purpose much different from this examination of gambling, a fortuitous circumstance which might minimize bias. Unfortunately, I did not have sufficient research funds to hire an independent coder so that independent ratings of the variables could be compared with mine.

A list of the societies in the sample, along with the codings of the different variables, can be found in the appendices. Further details may be obtained from the author.

E. The Calculations

The coding of the presence of gambling was carried out in two different ways. In one coding I set up three dummy variables, where 0 = gambling not reported; 1 = gambling reported but not an important economic activity; and 2 = gambling reported as important economic activity. "Importance" represented my subjective judgment on the relative value of the goods changing hands in such activities as a percentage of total societal income. Another type of gambling variable was also examined (where the dummy variable had only two values, denoting whether gambling was present or not), but the explanatory power of these regressions was somewhat less, and the experiments were abandoned. Several alternative measures of the level of economic development were also investigated and dropped because of lower explanatory power than the variable chosen.

---

[2] The codings on games of chance come from the *Ethnographic Atlas*, 1972 card version, obtained from the Ethnographic Atlas project of the Department of Anthropology, University of Pittsburgh. Of the 43 societies for which their and my codings could be compared, we agree on the presence or absence of gambling in 37 (86 percent) of these. There was no systematic bias shown in the error pattern of the six societies where our codings differed.

The dependent variable plus the 10 proposed independent variables are presented in table 1 below, along with their correlations. It should be noted that there is a certain amount of multicollinearity among the explanatory variables, and therefore we must adopt special procedures. The most simple is to select subsamples of variables and determine which groups lead to significant regression coefficients and the highest coefficients of determination, and this procedure was followed. The risk-loving hypothesis and the contact-with-the-West hypothesis quickly fell by the wayside. The variables connected with individualism (i.e., presence of lineages and the type of family) also did not seem very important and were dropped. The hardness-of-environment variable and the nomadic-herding variables are sufficiently correlated to each other that only one must be used in the statistical analysis, the latter was chosen because there was less subjectivity in its coding and also because it yielded somewhat higher coefficients of determination, a fortunate congruence of events.

In the final rounds of testing, five independent variables remained. The major difficulty arose in deciding whether the economic-development variable or the social-inequality variable should be retained in the regression. If we calculate a regression with all five independent variables, the coefficient for the development variable is not statistically significant, while the coefficient for the inequality variable is. Further, the coefficient of determination for the regression if the development variable is dropped is somewhat higher than if the social-inequality variable is dropped. The final regression can be shown thus

$$PG = -0.2811 + \underset{(.1514)}{1.3302}\, LNA + \underset{(.1256)}{0.4151}\, DCM$$

$$+ \underset{(.0746)}{0.2372}\, SI - \underset{(.1961)}{0.4299}\, NH \qquad R^2 = .6757 \quad n = 60$$

where PG = presence of gambling (0 = not present; 1 = present but unimportant or present and importance unascertainable, 2 = present and an important redistributive mode); LNA = location in North America

TABLE 1

Correlation Matrix of Gambling and Various Proposed Explanatory Variables

	Dependent Variable	Independent Variables									
	1*D*	1	2	3	4	5	6	7	8	9	10
Dependent variable:											
1*D*. Presence of gambling	1.00										
Independent variables:											
1. Location in North America	0.68	1.00									
2. Social inequality	0.27	−0.11	1.00								
3. Nomadic herding	−0.25	−0.16	0.11	1.00							
4. Presence of domestic commercial money	0.35	1.01	0.38	0.13	1.00						
5. Hardness of environment	−0.01	0.19	−0.05	0.68	−0.07	1.00					
6. Level of economic development	0.24	−0.16	0.74	−0.04	0.57	−0.17	1.00				
7. Hunting/fishing for 50% or more of total food supply	0.12	0.32	−0.24	−0.13	−0.17	0.16	−0.38	1.00			
8. Type of family (nuclear or extended)	0.07	−0.07	0.05	−0.14	0.17	−0.01	0.12	−0.16	1.00		
9. Contact with West	0.03	0.10	0.11	0.43	0.27	0.37	0.17	−0.13	−0.04	1.00	
10. Presence of lineages	0.03	0.20	−0.23	0.18	−0.11	0.37	−0.34	0.21	0.06	0.15	1.00

(excluding Central America and the Caribbean islands) (1 = yes; 0 = no); *DCM* = presence of a domestic commercial money (1 = yes; 0 = no); *SI* = presence of socioeconomic inequality (a five-point scale: 1 = little or no socioeconomic inequality; 5 = considerable socioeconomic inequality); *NH* = society nomadic or seminomadic and more than half of its food supply coming from animal husbandry (1 = yes; 0 = no); R^2 = coefficient of determination; n = size of sample, and numbers in parentheses = standard errors [3]

Several technical aspects of the regression deserve brief mention. First, this is the combination of four variables which has the highest coefficient of determination of any of the combinations which I tried. Second, the calculated regression coefficients are quite robust; when one of the variables is dropped from the calculations, there is not a great deal of change in the calculated slope coefficients of the other variables.

The most important feature of the regression for this discussion is that the calculated coefficient for the socioeconomic-inequality variable is statistically significant, thus confirming a deduction from the Friedman-Savage utility analysis The calculated coefficient for the nomadic herding variable is also statistically significant (at the 05 level) and confirms another aspect of the Friedman-Savage discussion. The significance of the coefficient for the money variable confirms a commonsense economic proposition which rests on intuitive notions about the implications of liquidity The only variable reflecting considerations discussed in the anthropology literature is the location variable, which shows, I believe, that the presence of gambling depends not only upon the presence of money or social inequality or the absence of nomadic herding but also upon some sort of diffusion phenomenon which remains to be explored

In short, these results suggest that deductions from economic theory about gambling serve us well in examining economic systems much different from the industrialized, capitalist economies to which our theories refer Further, we can derive from the strangely shaped Friedman-

---

[3] A problem in interpreting the statistical significance of the coefficients arises in the following manner· the presence of cultural diffusion of gambling among the 11 North American societies lowers the number of independent cases, and therefore our degrees of freedom are lowered It is noteworthy, however, that the calculated coefficients of most of the variables in the regressions continue to have high *t*-scores, even when the variable representing the location factor is removed For this reason, I do not believe that this problem of loss of degrees of freedom is serious Additional statistical problems arise because the dependent variable is a dummy variable, which means that the estimators are unbiased but not efficient (Kmenta 1971, p 427), and, further, because the error term is not normally distributed, the estimators of the constant coefficients are not normally distributed, and thus the classic tests of significance do not directly apply However, since it was the size of the coefficient of determination which provided the key test for the choice of the combination of variables, rather than the *t*-test, I do not believe that investment in more sophisticated estimation techniques for the coefficients of the independent variables would have a very high payoff (especially since examination of the regression residuals did not suggest that the heteroscedasticity problem was very severe)

Savage utility function several propositions which receive impressive empirical validation from our sample of 60 primitive and peasant economies Although the presence of gambling among societies in the sample cannot be explained by exclusive reliance on propositions derived from the Friedman-Savage analysis, they certainly help us explain an important part of the variance.

Appendix A

Societies in Sample

The listing below gives the societies in the sample, alternative names, locations, and dates to which the data apply.

1. *Alor*: Atimelang village, Alor, Lesser Sunda Islands, Indonesia; 1900 (before extensive contact with Chinese or Dutch).

2. *Amhara*. Gondar district, Ethiopia, 1900 (before extensive reforms and changes).

3 *Ao Naga*: Assam valley, India, 1885 (shortly before British conquest).

4 *Aweikóma*: Dalbergia, Santa Catarina, Brazil, 1913 (before "pacification").

5 *Azande* [Zande, Niam-Niam] primarily in northeast Zaïre, but also Sudan and the Central African Republic, 1905 (before European administration)

6 *Aztec* [Tenochca]: central Mexico; 1519 (shortly before arrival of Cortes)

7. *Bassen* [Khamseh] Fars, Iran, 1955 (shortly before Barth's fieldwork).

8. *Batak* [Toba-Batak]. Sumatra, 1880 (shortly before intensive missionary work).

9 *Bhil* area around Rajpipla, Gujarat, India; 1935 (shortly before the fieldwork of Koppers)

10. *Bribri* [Talamanca] Costa Rica, 1866 (before strong Western influence).

11 *Callinago* [Island Carib] Dominica, Lesser Antilles; 1640 (at the time of early French visits).

12. *China* [Kwangtung province peasant] Nanching village, Kwangtung province, China, 1930 (before modern road put in)

13. *Comanche* Texas and Oklahoma, U S A., 1850 (before "pacification").

14. *Copper Eskimo* Coronation Gulf, Northwest Territories, Canada, 1910 (before extensive Western contact)

15. *Dogon* [Habe, Kado] Sanga region, Mali, 1900 (before extensive Western contact).

16. *Fiji*. Vanua Levu, Fiji, 1820 (before extensive Western contact).

17 *Fon* [Dahomey] Dahomey, 1860 (before end of slave trade and beginning of extensive Western influence)

18. *Ganda* [Baganda]: Uganda; 1855 (before extensive Western contact).

19 *Gheg*. Albania, 1912 (at year of independence from Turkey).

20. *Havasupai*: Arizona, U S.A , 1890 (before extensive Western contact).

21. *Inca*: Peru, Bolivia, and Ecuador; 1532 (shortly before Spanish conquest).

22. *Iroquois:* New York, U.S.A.; 1700 (before extensive Western influence).

23. *Khalkha Mongol:* Mongolia; 1910 (before gaining independence).

24. *Koryak* [Reindeer Koryak, Chavchuven]: Kamchatka peninsula and adjacent areas; 1901 (at time of fieldwork of Jochelson).

25. *!Kung Bushmen:* Southwest Africa and Botswana; 1950 (shortly before fieldwork of the Marshall family).

26. *Kwakiutl* [Southern Kwakiutl]: Vancouver Island, British Columbia, Canada; 1860 (before extensive contact with the West).

27. *Lapp* [Mountain Lapp]: Käresuando parish, Sweden; 1950 (shortly before fieldwork of Whitaker and Pehrson).

28. *Lepcha* [Rong]: Sikkim; 1870 (before extensive Western contacts).

29. *Manóbo:* Agúsan valley, East Mindanáo, Philippines; 1905 (at time of fieldwork of Garvan).

30. *Maori:* north part of North Island, New Zealand; 1800 (before extensive white contact).

31. *Mundurucú:* Cabrúa village, Pará, Brazil, 1952 (at time of fieldwork of Murphy).

32. *Murngin* [Wulamba]. east Arnhem land, Northern Territory, Australia; 1926 (at time of fieldwork of Warner).

33 *Naskapi:* north Labrador peninsula, Canada, 1880 (shortly before fieldwork of Turner).

34 *Navajo:* Arizona and New Mexico, U.S.A., 1860 (shortly before "pacification").

35. *Nuer:* Ā Ālī A-Nīl, Sudan; 1920 (before "pacification").

36 *Nyakyusa* [Banyakyusa, Wanyakyusa] southern highlands, Tanzania; 1890 (shortly before missionary arrival).

37. *Omaha:* Nebraska, U S A., 1850 (before extensive Western contact)

38 *Pomo* [Eastern Pomo] Clear Lake, California, U S A., 1847 (shortly before California gold rush).

39. *Rif* [Aith Waryaghar Rif]. Morocco, 1920 (before major changes)

40. *Rwala* [Rwala Bedouin]. Syria, Jordan, and north Arabia, 1905 (at beginning of fieldwork of Musil)

41. *Semang* [Orang-Utan, Sakai] northwest Malaysia, 1924 (at time of first fieldwork by Schebesta)

42 *Serbia:* Sumadija region, Serbia, Yugoslavia, 1910 (before Serbian-Turkish war)

43 *Shavante* [Akwe-Shavante, Chavante, Xavante] central plateau, east Mato Grosso, Brazil, 1950 (before extensive Western contact).

44. *Siane:* Goroka subdistrict, eastern highlands, New Guinea, 1944 (before Western contact).

45. *Sirionó* eastern Bolivia, 1940 (at time of fieldwork by Holmberg).

46. *Suku* [BaSuku, Pindi] Kwango district, Zaïre, 1915 (before extensive contact with West)

47. *Tanala* [Menabe Tanala] Malagasy; 1890 (before extensive contact with the West).

48 *Thonga* [BaThonga] Mozambique, 1870 (before beginning of fieldwork of Junod)

49 *Tikopia*. Santa Cruz Island, Solomon Islands; 1928 (at beginning of Firth's fieldwork)

50. *Tiv* [Munshi] eastern central Nigeria, 1900 (before extensive Western contact).

51. *Toba* [Eastern Toba, Toba-Pilagá, Pitilagá, Takshik, Komlik] Chaco and Formosa provinces, Argentina, 1860 (before lumber mills and military campaigns).

52. *Toda:* Nilgiri hills, Madras, India, 1901 (at time of Rivers's fieldwork).

53 *Trobriand:* Kiriwina district, Boyowa Island, Trobriand Islands, 1914 (at time of Malinowski's fieldwork).

54. *Truk:* Romonum Island, Truk Islands, Caroline Islands, 1895 (before extensive Western or Japanese contact).

Appendix B

TABLE B1

Sample Societies and Coded Variables

Society	1	2	3	4	5	6	7	8	9	10	11
Alor	1	0	0	1	38	0	0	0	2.0	1	0
Amhara	1	0	0	1	58	1	0	0	4.0	1	0
Ao Naga	1	0	0	1	33	1	0	0	2.2	1	0
Aweikóma	0	0	0	1	6	0	0	1	1.4	0	0
Azande	0	0	0	1	41	0	0	0	2.0	0	0
Aztec	2	0	0	1	56	1	0	0	4.5	1	1
Basseri	*0*	*0*	*1*	*1*	25.5	*1*	*1*	*0*	2.5	*0*	*0*
Batak	2	0	0	1	34	1	0	0	3.2	1	0
Bhil	0	0	0	1	47	1	1	0	2.5	1	0
Bribri	0	0	0	2	24	1	0	0	1.5	1	0
Callinago	0	0	0	1	27	0	0	1	1.8	0	0
China	2	0	0	1	60	1	0	0	3.3	1	0
Comanche	2	0	0	1	16	1	0	1	2.3	0	1
Copper Eskimo	1	1	0	1	9	0	0	1	1.4	0	1
Dogon	1	0	0	1	48	1	0	0	2.7	1	0
Fiji	0	0	0	1	45	0	0	0	2.3	1	0
Fon	1	0	0	2	51	1	1	0	4.8	1	0
Ganda	0	0	0	1	49	1	0	0	3.3	1	0
Gheg	1	0	0	2	50	1	0	0	2.0	1	0
Havasupai	2	0	0	1	14.0	0	0	0	1.3	0	1
Inca	1	0	0	1	52	0	0	0	4.7	0	0
Iroquois	2	0	0	1	28	0	0	0	1.8	1	1
Khalka Mongol	1	1	1	1	39	1	1	0	3.3	0	0
Koryak	0	1	1	1	12	0	1	0	2.2	0	0
!Kung Bushmen	0	0	0	1	4.0	0	0	0	1.2	0	0
Kwakiutl	2	0	0	1	29	1	0	1	2.5	1	1
Lapp	0	1	1	1	20.5	1	1	0	1.8	0	0
Lepcha	0	0	0	1	40	0	0	0	2.3	1	0
Manóbo	0	0	0	1	22	0	0	0	1.8	0	0
Maori	0	0	0	1	42	0	0	0	3.2	0	0
Mundurucú	0	0	0	1	11	0	1	0	1.3	1	0
Murngin	0	0	0	1	4.0	0	0	0	1.2	1	0
Naskapi	1	1	0	1	7	0	1	1	1.1	0	1
Navajo	2	1	0	2	23	0	1	0	2.0	1	1
Nuer	0	0	0	1	19	0	0	0	1.9	1	0
Nyakyusa	0	0	0	1	31	0	0	0	2.4	0	0
Omaha	2	0	0	1	17	0	1	0	2.3	1	1
Pomo	2	0	0	1	18	1	0	0	1.7	0	1
Rif	0	0	0	1	55	1	1	0	2.5	1	0
Rwala	0	1	1	1	37	1	1	0	2.4	1	0
Semang	0	0	0	1	2	0	0	0	1.2	0	0
Serbia	1	0	0	1	57	1	1	0	2.8	1	0
Shavante	0	0	0	1	8	0	0	0	1.0	1	0
Siane	0	0	0	1	10	1	0	0	1.5	1	0
Sirionó	0	0	0	1	1	0	0	1	1.3	0	0
Suku	1	0	0	1	44	1	0	0	2.0	1	0
Tanala	1	0	0	2	43	1	0	0	2.3	1	0
Thonga	0	0	0	2	35	0	0	0	2.4	1	0
Tikopia	0	0	0	1	30	0	0	0	2.5	1	0
Tiv	1	0	0	2	46	1	0	0	2.3	1	0
Toba	1	0	0	1	14.0	0	0	1	2.2	0	0
Toda	0	0	0	1	20.5	0	0	0	2.2	1	0
Trobriand	0	0	0	1	25.5	0	0	0	2.3	1	0
Truk	1	0	0	2	32	1	0	0	2.0	1	0
Tuareg	0	1	1	1	36	0	0	0	3.0	1	0
Turkey	1	0	0	1	59	1	1	0	2.8	0	0
Warao	0	0	0	2	14.0	0	0	0	2.0	1	0
Wolof	1	0	0	1	53	1	1	0	2.5	1	0
Yahgan	0	1	0	1	4.0	0	0	0	1.0	0	0
Yaqui	1	0	0	1	54	1	1	0	1.8	0	1

55. *Tuareg* [Ahaggar Tuareg]: Ahaggar area, Algeria; 1898 (before French military occupation).

56. *Turkey* [Turk]: Sakaltutan and Elbasi villages, central Anatolia; 1950 (at beginning of Stirling's fieldwork).

57. *Warao* [Warrau, Guaraunos]: Orinoco delta, Venezuela; 1920 (before introduction of Western agricultural techniques).

58. *Wolof* [Ouolof]: Senegal and Gambia; 1875 (before conquest by French).

59. *Yahgan* [Yamana]: Tierra del Fuego, Chile; 1850 (before missionaries).

60. *Yaqui:* Sonora, Mexico, and Arizona, U.S.A.; 1930.

Description of Data Series

1. Presence of gambling: 0 = none reported; 1 = some reported but either not important or importance unascertainable; 2 = gambling reported and an important redistributive mode

2. Hardness of the environment 1 = hard environment (e g., arctic, semi-arctic, or desert); 0 = not hard environment.

3. Nomadic herding · 1 = animal husbandry accounts for more than 50 percent of food supply, and society is either nomadic or seminomadic, 0 = one or both of these conditions are not filled. (Further details on the coding of this variable may be found in Pryor, forthcoming)

4. *Unit of observation · 1 = nuclear family (monogamous, polygynous, poly-*androus) or small extended family, 2 = extended family (Further details on this variable may be found in Pryor, forthcoming)

5 Rank order of level of economic development 1 = rank of lowest level; 60 = rank of highest level. The decimal points indicate tie-scores. (Ranking follows an approach used by Carneiro 1973, further details may be found in Pryor, forthcoming).

6 Presence of a domestic money used for commercial purpose · 1 = yes; 0 = no. (Excluded from consideration were noncommercial moneys and moneys used exclusively in external trade)

7. Contact with the West 0 = no contact, occasional contact, or light missionary activity, 1 = considerable contact, heavy missionary work, or considerable contact with Western government officials, colonists, visitors, or traders.

8. Hunting and/or fishing accounting for 50 percent or more of total food supply · 1 = yes, 0 = no.

9 Socioeconomic inequality, computed as an average of six five-point scales indicating inequality of total income, inequality of consumption, inequality of ownership of property, inequality of menial work carried out, inequality of total work carried out, and inequality in the holding of technological knowledge: 1 = least inequality; 5 = greatest inequality. (These scales were subjectively coded)

10 Important presence of extensive lineages, defined in terms of group economic activities (lineages include patrilineal, matrilineal, duolineal, and ambilineal groups): 1 = yes, 0 = no.

11. Presence in North America (excluding Central American and Caribbean islands) · 1 = yes, 0 = no.

References

Boserup, Ester *The Conditions of Agricultural Growth.* London · Allen & Unwin, 1965.

Carneiro, Robert L. "Scale Analysis, Evolutionary Sequence, and the Rating of

Cultures." In *A Handbook of Method in Cultural Anthropology*, edited by Raoul Naroll and Ronald Cohen. New York: Columbia Univ. Press, 1973.

Collins, Paul W. "Functional Analyses in the Symposium 'Man, Culture, and Animals'" In *Man, Culture and Animals*, edited by A. Leeds and A. P. Vayda. Publication no 78 of American Assoc. Advancement Sci. Washington, 1965.

Friedman, Milton, and Savage, L. J. "The Utility Analysis of Choices Involving Risk." *J.P.E* 56, no. 4 (August 1948): 279–304.

Kmenta, Jan. *Elements of Econometrics*. New York: Macmillan, 1971.

Lee, Richard "What Hunters Do for a Living, or, How to Make out on Scarce Resources" In *Man the Hunter*, edited by Richard B. Lee and Irven DeVore. Chicago Aldine, 1968.

Lowie, Robert H. *An Introduction to Cultural Anthropology* New York Farrar & Rinehart, 1934

Markowitz, Harry "The Utility of Wealth" *J P E* 60, no 2 (April 1952): 151–58

Murdock, George Peter. "World Ethnographic Sample." *American Anthropologist* 59 (August 1957) 664–87

Murdock, George Peter, and Provost, Caterina "Measurement of Cultural Complexity" *Ethnology* 13 (October 1973). 379–99.

Pryor, Frederic L. *The Origins of the Economy A Comparative Study of Distribution in Primitive and Peasant Societies* Forthcoming.

Roberts, John M, and Sutton-Smith, Brian "Cross-cultural Correlates of Games of Chance" *Behavior Science Notes* 1 (March 1966) 131–44

Sahlins, Marshall "The Original Affluent Society" In *Stone Age Economics* Chicago Aldine, 1972

Fertility Interactions and Modernization Turning Points

Paul R Gregory
University of Houston
J M Campbell, Jr
University of Chicago and Economic Analysts, Inc

I. Introduction

In this paper we estimate an interaction model of fertility and economic development which directly incorporates into the fertility decision the taste changes and variations in fertility control associated with modernization In this manner, the impact of changes in attitudinal and control factors on underlying economic relationships can be studied and the pattern of elasticity response elicited, including a specific test of the variation of the child-quantity income elasticity as the country develops economically The ordinary least-squares (OLS) interaction results are then checked against a two-stage least-squares (2SLS) model in which the impact of fertility upon per capita income is endogenized, and we show that the basic OLS results are not seriously affected by simultaneous-equation bias

Our interaction approach seeks to modify the existing empirical literature[1] by incorporating directly the impact of "modernization" on the pattern of income and substitution fertility elasticities and thereby creating a dynamic model of fertility and economic development. The policy implications of our switch from static to dynamic modeling are that fertility responses to a given change in a particular explanatory variable will vary with the level of modernization, and that population policy must, therefore, be geared to the country's stages of modernization. We hope that our results shed light on the debates concerning the linkage between economic development and fertility at different stages of

---

The authors would like to thank the late Stephen Enke and also Mike Keeley for their comments on an earlier draft We are grateful to Sam Peltzman and an anonymous referee of this *Journal* for their helpful suggestions The usual caveat applies

[1] A review of the literature in the context of the controversy over economic development and fertility is found in our forthcoming paper (see Gregory and Campbell 1976)

[*Journal of Political Economy*, 1976, vol 84, no 4, pt 1]
© 1976 by The University of Chicago All rights reserved.

economic maturity[2] and on the relative importance of income, price, attitudinal, and demographic variables in explaining secular fertility decline.[3] Our major finding is that income and substitution elasticities do change in a rather dramatic fashion during the course of economic development, and that many of the competing propositions in the economic theory of fertility can be reconciled in terms of these elasticity patterns

Our fertility interaction models are fitted to a combined time-series cross section of 18 Latin American countries (1950 and 1960) The sample is limited to Latin America in order to achieve the homogeneity essential to fertility analysis,[4] and a multiperiod sample is used to better capture secular trends in the fertility–economic development relationship [5] A list of variables and data sources are provided in the Appendix All data are drawn from readily available United Nation's publications and follow UN definitions We do not pretend that our data measure perfectly the price, income, and modernization variables called for by the theory discussed below Rather, ad hoc proxies which are commonly used as approximations of theoretical variables are included, and our results must be interpreted with this important reservation in mind Furthermore, it would have been desirable to deal with cross sections covering a longer time span in view of the dynamic implications of our model, but data limitations prohibited this approach

II. An OLS Interaction Model of Fertility

We begin with a static model which assumes that the demand for births, as measured by the crude birthrate (CBR), is "explained" by income and substitution effects, holding the level of "modernization" constant Thus our ceteris paribus demand-for-births model is

$$\text{CBR} = f(\text{PY}, \text{FRP}, \text{IM}, \text{IL}, \text{UP}), \tag{1}$$

where

PY = per capita income (1960 U S dollars),
FPR = female labor participation rate,
IM = rate of infant mortality,
IL = illiteracy rate,
UP = percentage of urban population.

---

[2] For differing viewpoints, see Nelson (1956), Enke (1971), Hansen (1972), Venieris, Sebold, and Harper (1973), and Kelley (1974)

[3] For different opinions on this issue, consult Becker (1960), Coale (1967), Leibenstein (1973, 1974), and Griliches (1974)

[4] Simon (1974) shows that the various fertility elasticities become more clearly defined as the sample becomes more homogeneous

[5] Gregory and Griffin (1974), for example, demonstrate that secular trends are better elicited by multiperiod samples than by a single cross section

The rationale for including these various income and cost proxies has been discussed elsewhere, but a brief discussion seems warranted. We emphasize that we are applying at this point narrow economic interpretations (through income and substitution effects) to the variables in (1) to simplify the exposition. Noneconomic effects arising during the process of modernization are assumed to be captured by the modernization interaction functions discussed below, although we recognize that changes in economic and quasi-economic variables may also yield similar interaction effects (Becker and Lewis 1973, pp 278–85)

We use the crude birthrate (CBR) to measure fertility instead of a measure of completed fertility, because this measure may be more accurate for low-income countries where changes in the timing of births seem to be of much less consequence in accounting for changes in birthrates over time as the relationship between the stock and the flow tends to be fairly stable (Simon 1974, pp. 15–21) Earlier experiments with the general fertility rate (GFR) as an alternate dependent variable were dropped due to suspect age-distribution data We recognize that the CBR measure is by no means an ideal dependent variable, but the correlation between CBR and other fertility measures (the GFR, the child-woman ratio, etc) tends to be quite high,[6] and CBR should serve as an adequate fertility proxy

As far as the cost and income variables are concerned, per capita income (PY) is included because one would expect fertility decisions to change with income, holding tastes, fertility control, and fertility costs constant This income effect could range from positive to possibly insignificant, depending on the ease with which child quality may be substituted for the quantity of children, and may vary with income (Willis 1973, pp 34–41) The use of PY introduces the prospect of simultaneous-equation bias due to its negative correlation with fertility, but we show below that the impact of such bias is small. Infant mortality (IM) will be positively related to births as a proxy for the replacement demand for children where a highly inelastic demand for surviving offspring exists On the other hand, IM may be negatively related to births through the substitution effect, insofar as increases in child mortality may increase the cost of rearing a child to maturity The net effect of IM will therefore depend on which effect tends to dominate

The percent of urban population (UP) is anticipated to vary inversely with fertility, insofar as one would expect the costs of fertility (both the direct commodity costs and indirect opportunity costs of childbearing and raising) to be positively related to the level of urbanization Thus UP is postulated to act as a proxy for the price (cost) of children. Another cost variable is the female labor participation rate (FPR) The reasoning behind its inclusion in the static model is the expected trade-off between fertility and female labor force participation, because mothers must withdraw from the labor force to bear and rear (at least to nursery school age) their children (see Gregory, Campbell, and Cheng 1972; Kim et al. 1972) This withdrawal from the labor force will tend to generate costs in addition to the immediate loss of earned income in the form of lost experience and seniority (depreciation of human capital), which will lessen the wife's lifetime earning capacity (Willis 1973, pp. 22–25)

---

[6] This is the finding of Bogue and Palmore (1964) from an investigation of a larger intercountry sample.

Some further discussion of the FPR variable may be in order, for monetary measures of time value, an observed wage rate is usually used to measure fertility opportunity costs. Arguments for wage rates and against FPR generally rely on Mincer's theorizing that FPR is the inverse of the birthrate in static demand analysis and Willis's less accepted arguments for linearity between time value and observed wage rates (Mincer 1963; Willis 1973, pp 30–37) However, the wage-rate time-value approach also involves some recognized deficiencies For nonworking women, Gronau has shown that wage rates underestimate time value, while the time value of part-time female labor participants has also been shown to be incorrectly measured by wage rates in Willis's work (Gronau 1973, pp. 175–76, Willis 1973, pp 30–37). Obviously, the researcher is left with two imperfect opportunity cost measures

Given this dilemma, we have opted for the FPR measure in a study which concentrates on modernizing countries, because wage-rate data restricted to industrial workers are believed to be a poorer proxy for opportunity costs of a mother's time than in developed-country studies (Schultz 1969, pp 154–56) and because the wage rate's effect on fertility may be captured by the extension of equation (2) to account for the simultaneous relationship between fertility and income and by the inclusion of education Obviously, both measures are imperfect proxies For the benefit of those who are skeptical of proxies, it should be noted that econometric theory has developed some preliminary evidence which shows that the loss of statistical precision resulting from specification error, such as leaving out opportunity costs, exceeds that from inadequate proxies except when the proxy measurement error is very large (Wickens 1972)

Finally, the illiteracy rate (IL) is included as a general proxy for education of parents In the static model, education may directly affect fertility demand as a proxy for parents' wages insofar as male and female wages are positively related to the level of education. It should be noted that a net wage effect is being captured here, as the wife's wage will probably be negatively related to fertility (a price of time substitution effect), whereas the husband's wage rate would be positively related to fertility through the income effect. Third, education will tend to affect the wife's productivity in home activities and thus represent the shadow price of the wife's time in the home (Gronau 1973, pp. 178–80).

Again we stress the relatively crude nature of our empirical model, and that it does not necessarily capture pure income and substitution effects. The income variable may pick up some price-of-time substitution effects, the IM variable may capture stock adjustment and substitution effects, the IL variable may pick up a whole series of effects, and so on

We amend the basic demand model (1) by introducing in the model an interaction function, *g*, which incorporates the shifts in tastes, attitudes, and contraceptive efficiency which accompany the process of modernization We assume that "modernization" can be measured via its relationship to urbanization (UP) and education (IL), where IL will capture both taste changes and contraceptive efficiency We recognize that modernization cannot be fully captured by UP and IL alone, but this approach seems to represent a reasonable approximation given the crude nature of the other variables and the regional homogeneity of our sample The interaction function *g* is estimated in two variants $g_A(\mathrm{UP})$ and $g_B(\mathrm{UP, IL})$, where the complete model is

$$\mathrm{CBR} = f(\mathrm{PY, FPR, IM, IL, UP}) \cdot g_A(\mathrm{UP}), \tag{2a}$$

or

$$\mathrm{CBR} = f(\mathrm{PY, FPR, IM, IL, UP}) \quad g_B(\mathrm{UP, IL}) \tag{2b}$$

For purposes of empirical estimation, both *g* and *f* are assumed linear[7] to yield the interaction models reported in table 1 The most critical feature of the interaction model is that it allows the various income and cost effects of the demand model to vary with modernization as the demand factors interact with general modernization factors

Why would one expect significant interactions of this type? With modernization, significant changes in the nature of income and cost relationships may occur For example, as urbanization proceeds, the "producer good" nature of children diminishes, thus changing the basic fertility cost calculus Moreover, the opportunity cost of children in terms of foregone female earnings should rise as formal ties to the industrial labor force must be broken (and human capital depreciated) to bear and rear children, whereas such ties are more likely to be maintained in a rural setting Improved fertility control in the course of modernization will cause reductions in excess fertility—a phenomenon captured by an

---

[7] A similar interaction has been applied to U S time-series data by Wachter (1974) We also experimented with nonlinear *g* functions. The introduction of such nonlinearities failed to improve the results described in this paper

TABLE 1

OLS INTERACTION MODEL, LATIN AMERICAN FERTILITY, COMBINED TIME-SERIES CROSS SECTIONS (1950 AND 1960)

MODEL	REGRESSION COEFFICIENTS WITH RESPECT TO														
	Constant	PY	UP	FPR	IM	IL	IL UP	PY UP	FPR UP	IM UP	PY · IL	FPR IL	IM IL	IL²	UP²
1	42.1	0.22 (2.2)	−0.40 (2.5)	−0.35 (2.2)	0.09 (1.9)	0.10 (1.2)									
2a	−33.4 (1.0)	1.03 (2.4)	1.58 (1.4)	1.05 (2.1)	0.16 (1.0)	0.48 (1.9)	−0.008 (1.5)	−0.02 (1.8)	−0.04 (2.8)	−0.002 (0.7)					−0.003 (0.3)
2b	−81.3 (1.3)	1.42 (1.9)	2.23 (1.1)	0.53 (0.3)	0.57 (1.5)	1.26 (1.8)	−0.009 (0.6)	−0.02 (1.9)	−0.03 (1.3)	−0.007 (1.2)	−0.005 (0.5)	0.005 (0.3)	−0.006 (1.3)	−0.002 (0.3)	−0.005 (0.3)

NOTE.—t-values in parentheses, for model (1), $\bar{R}^2$ = 40, SE = 7.0, F = 5.8, for model (2a), $\bar{R}^2$ = 58, SE = 5.9, F = 5.7, for model (2b), $\bar{R}^2$ = 55, SE = 6.1, F = 6.1

interaction model (Wachter 1974, pp. 5–10). Finally, the pure income effect may vary with income (Becker and Lewis 1973, pp. 280–83), and, insofar as our per capita income measure also captures substitution effects (as a proxy for wages and the price of time), the measured income effect may vary as the relative weights of income and substitution effects change during modernization.

Equation (2) is estimated using OLS, and the results are recorded in table 1 Model (2a), based on the interaction function g_A(UP), provided the "best" results in terms of coefficient significance, fit, and standard error. Model (2b), employing the more complex interaction function g_B(UP, IL), yields similar results For reference purposes, the noninteraction static demand model results are also recorded in table 1. Space constraints prevent our dwelling on these results, for our principal concern is the variability of elasticities during modernization It is perhaps sufficient to note that our OLS results provide support for the interaction model in terms of significance of interaction coefficients, especially in equation (2a) Deserving of special attention are the strong interactions of PY and FPR with UP Comparison of the two interaction models (2a and 2b) with the noninteraction results (1) indicates significant reductions in unexplained variance and improvements in standard errors as a consequence of introducing the interaction terms the adjusted coefficient of multiple determination is improved by almost 50 percent, and the standard error of the estimate declines by almost 20 percent

III. Modernization Turning Points: An Investigation of Elasticity Variability

In this section, we concentrate on the variability of income and substitution elasticities in response to changes in modernization. As noted previously, "modernization" is incorporated in the model in the form of the interaction functions, g_A(UP) and g_B(UP, IL) First, we calculate the elasticities of CBR with respect to the explanatory variables in the model (PY, FPR, IM, IL, and UP), evaluated at the sample means Second, we calculate "modernization turning points" (MTPs) from the partial derivatives of CBR with respect to PY, FPR, IM, IL, and UP. This is done by determining the values of UP (our principal modernization proxy) at which the various elasticities change sign, if at all. In this manner the MTPs provide crude summary measures of the pattern of elasticity change in the course of modernization and some rather stark detail on how societies in different stages of modernization respond differently to price and income changes The mean elasticities and MTPs are recorded in table 2. The figures in parentheses are derived from g_B(UP, IL), and the nonparenthesized figures are derived from the "better" g_A(UP) interaction results (table 1) The g_A elasticities and

TABLE 2

MEAN ELASTICITIES, MODERNIZATION TURNING POINTS (MTPs), AND ELASTICITY VARIATION FOR LATIN AMERICAN FERTILITY (CBR)

INDEPENDENT VARIABLE	MEAN CBR ELASTICITIES	VALUE OF UP AT WHICH ELASTICITY IS					ELASTICITY
		0 10	0 05	0(MTP)	−0 05	−0 10	
PY . .	0 30 (0 35)	60	64	69(68)	72	76	−Above 69(68)
FPR	−0 23 (−0 28)	21	25	28(22)	31	35	+Below 28(22)
IM	0 16 (0 07)	58	69	80(47)	90	100	−Above 80(47)
IL	0 15 (0 11)	48	54	60(54)	66	72	−Above 60(54)
UP	−0 22 (−0 26)	Negative	3	10(20)	17	24	+Below 10(20)

NOTE.—The figures not in parentheses are derived from table 1, model (2a), the figures in parentheses are derived from table 1, model (2b)

MTPs are stressed in the discussion below.[8] The mean CBR elasticities are to be interpreted as the percent change in CBR brought about by a 1-percent change in the independent variable evaluated at the sample means. The interaction impact of modernization on the various income and substitution effects is directly incorporated into these elasticities (as evaluated at $\overline{UP}$) insofar as the elasticities are functions of UP. Thus, the mean CBR elasticities relate the response of CBR to changes in the various explanatory variables at an average level of "modernization."

The mean CBR elasticities provide no surprises. The PY income elasticity is a positive 0.30; the UP, FPR, and IL substitution elasticities have the expected signs (−0.22, −0.23, and 0.15, respectively), and the IM elasticity is positive (0.16), indicating a domination of the replacement demand effect over the substitution effect. Looking at the MTPs in table 2, we find considerable variability in elasticities. The PY elasticity changes from positive to negative at a relatively high UP value (UP = 69 and above)—a finding which seems to indicate a lessening of the income effect on child quantity as high levels of modernization are achieved, possibly due to quality substitutions (Wachter 1974, pp. 2–6) or to shifts in tastes toward high-status consumption commodities (Leibenstein 1973, pp. 5–9). Most important, this finding illustrates why previous fertility studies have found both positive and negative income effects on fertility, as it suggests that the positive income effect on fertility at low levels of economic development gradually becomes negative in the course of modernization.

The FPR elasticity changes from positive to negative at relatively low levels of modernization (UP = 28 and above). This elasticity shift is explained by the fact that as urbanization proceeds, the movement of employed women from the agricultural sector raises opportunity costs because mothers must withdraw from the industrial labor force to bear and rear children. The IM elasticity changes sign only at quite high levels of urbanization (UP = 60 and above). This finding seems to indicate that ceteris paribus the replacement effect of infant mortality tends to dominate its cost impact over the course of modernization, indicating that children are primarily producer goods in less developed societies. The IL elasticity is positive up to UP = 60, after which it becomes negative. A possible (but unlikely, in view of the weak income effect at high income levels) explanation for this finding is that at higher levels of modernization, reductions in IL will capture primarily income effects; at lower levels of modernization, primarily substitution effects are captured. A more likely explanation is the variability of fertility control during modernization and the fact that modernization and contraceptive technology would be expected to be positively related (Wachter 1974, pp. 9–14). Finally, we note that the UP elasticity is negative throughout almost all stages of modernization, indicating its consistent negative impact on CBR through substitution effects and taste shifts.[9]

---

[8] Comparison of the g_A and g_B MTP sets (table 2) indicates relatively close accord with the exception of the IM (80 vs. 47) MTP. We also experimented with a third interaction function g_C(IL), where the turning points are defined by IL. This third approach was rejected because of generally poor regression results (only one significant primary coefficient and one significant interaction term, marginally lower standard errors, etc.) We should note that the g_C turning points are generally consistent with the table 2 MTPs. The FPR and IM turning points occur at low (IL = 72) and below average (IL = 57) modernization, and the UP elasticity remains negative over the entire modernization range. The PY and IL turning points, however, are unlike their table 2 counterparts, suggesting positive elasticities over the modernization range. We should note, however, that the PY and IL turning points are based on quite insignificant primary and interaction coefficients. We have not included PY in the interaction function because "modernization" is supposed to capture only taste changes and fertility control effects, which would be only indirectly linked to PY, while directly linked to UP and IL. Anyway, the addition of PY to the g_B interaction term would have only a minor impact on the MTPs, as UP and IL account for most of the variation in PY.

IV. An Investigation of Elasticity Ranges and Application to Latin America

The foregoing discussion attempts to pinpoint critical fertility elasticity turning points during the process of modernization. However, further calculations are required to determine over what UP ranges the respective elasticities are "small" and thus exert relatively unimportant impacts on fertility. In table 2 we calculate (using the g_A[UP] interaction function) the respective UP values which yield elasticities of between +0.10 and −0.10 for the respective independent variables. In this manner, we provide crude calculations of modernization ranges over which the elasticity is small (adopting an arbitrary standard of $<|0.10|$ for the various elasticities) and likely to appear as insignificant in regression models based on samples representing different states of modernization.

For example, the income (PY) elasticity is $<|0.10|$ for urbanization ratios of between 60 and 76 percent, which suggests, when combined with the mean elasticity of 0.30, that significant positive income effects will be obtained only at low rates of urbanization. The FPR elasticity, on the other hand, is $<|0.10|$ for UP values between 21 and 35 percent, suggesting significant FPR effects at relatively high UP proportions—over 50 percent, for example. The IM elasticity is $<|0.10|$ for all urban proportions over 58 percent; therefore, the IM impact seems to be relevant only at relatively low UP values. Finally, the UP elasticity is

---

[9] To test the extent of simultaneous-equation bias, we endogenized PY as a function of estimated CBR, per capita capital, labor, and labor quality inputs. Our two-stage least-squares estimations indicated that the only important change is the increase in the PY turning point—a direct consequence of the endogenization of PY.

TABLE 3

APPROXIMATE DATING OF MODERNIZATION TURNING POINTS (MTPs)
IN LATIN AMERICA

Independent Variable	Approximate Dating of MTP
PY	1990
FPR	Mid-1930s
IM	2000
IL	Late 1970s
UP	1900

$<|0.10|$ for almost all observable UP ranges (that is, UP values of 24 percent and above), which indicates the consistency of urbanization's negative impact on CBR.

As a final exercise, we apply the MTPs and the elasticity ranges (table 2) to the Latin American mean UP values in order to pinpoint Latin America's present and projected future location in this elasticity pattern. Of particular interest are the approximate dates when Latin America passed through or is projected to pass through the various modernization turning points. These dates are arrived at in a very approximate manner by applying average Latin American urbanization rates at various points in time to the MTPs in table 2 and are given in table 3. We stress that our results should be regarded as only illustrative of rough orders of magnitude because of the crude underlying urbanization information. On this basis, we calculate the UP and FPR turning points were probably attained in the 1900s and 1930s, respectively. On the other hand, the IL and PY turning points have yet to be attained and should be reached in the 1970s and 1990s, respectively. Finally, the IM turning point is calculated as occurring around the turn of the century. We again emphasize that the turning point represents a point in time when the impact of the variable on fertility is negligible, and, according to table 2, a considerable time period may be required before a significant fertility impact will be experienced.

V. Conclusions and Summary

A nonlinear interaction model of fertility is postulated to allow income and cost coefficients to vary explicitly with "modernization." The model is fitted using OLS to a combined time-series/cross-section sample of Latin American countries (1950 and 1960), and it is shown that fertility income and substitution effects do vary considerably during the course of modernization. In fact, the majority of income and substitution elasticities change signs during modernization, and these results are shown to hold in a simultaneous-equation setting in which the reverse effect of fertility on income is endogenized. The variation in elasticities supports our position that the competing views regarding the relationship between fertility and modernization can be partially reconciled through the recognition of this variability during the course of modernization. The basic policy conclusion emerging from this study is that caution must be exercised in applying the fertility experiences of one set of countries to another set of countries at a different level of economic development.

As far as the prospects for significant fertility reduction during the early stages of modernization are concerned, our findings are mixed. At low modernization levels, the income effect is indeed positive, while many of the retarding substitution effects (such as the FPR effect) do not begin to operate in the expected manner until relatively higher levels of modernization are attained. On the other hand, other factors such as infant mortality and education tend to have their greatest impact at low modernization levels, but their impact is quantitatively smaller than the income effect at low urban proportions.

Appendix

List of Variables and Sources of Data

1. Crude birth rate (CBR) is the number of registered births per 1,000 population. Source of data: *Demographic Yearbook*, various years, United Nations. The $\overline{CBR}$ = 38.3, CBR range is 18.6–51.6.

2. Per capita product (PY) is in 1960 constant U.S. dollars. Source of data: *Yearbook of National Accounts*, 1969, United Nations. The $\overline{PY}$ = \$302.5.

3. Infant mortality (IM) is the number of deaths of infants under 1 year of age per 1,000 live births. Source of data: *Demographic Yearbook*, various years, United Nations. The $\overline{IM}$ = 89.2, IM range is 49.1–170.0.

4. Percentage of urban population (UP) is the number of urban residents divided by total population. Source of data: *Demographic Yearbook*, various years, United Nations. The $\overline{UP}$ = 43.3, UP range is 23.8–82.1.

5. Female labor participation rate (FPR) is the percentage of the female population considered active in the labor force. Source of data: *Demographic Yearbook*, various years, United Nations. The $\overline{FPR}$ = 15.2, FPR range is 6.1–42.2.

6. Illiteracy rate (IL) is the number of people over 15 years of age who are illiterate divided by the total number of people ages 15 and above. Source of data: *Demographic Yearbook*, various years, United Nations. The $\overline{IL}$ = 39.9, IL range is 8.6–73.0.

References

Becker, G. "An Economic Analysis of Fertility." In *Demographic and Cultural Change in Developing Countries*. Princeton, N.J.: Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1960.

Becker, G., and Lewis, H. G. "On the Interaction between the Quality and Quantity of Children." *J.P.E.* 81, no. 2, suppl. (March/April 1973): S279–S288.

Bogue, D. J., and Palmore, J. A. "Some Empirical and Analytic Relations among Demographic Fertility Measures with Regression Models for Fertility Estimation." *Demography* 1 (1964): 15–24.

Coale, A. J. "Factors Associated with Development of Low Fertility· A Historic Summary." In *Proceedings of the World Population Conference, 1965* Vol 2 New York· United Nations, 1967

Enke, S. "Economic Consequences of Rapid Population Growth" *Econ. J* 48 (December 1971)· 61–69.

Gregory, P R , and Campbell, J M "Fertility and Economic Development—an Overview" In *Population, Policy and Economic Development*, edited by M Kelley New York Praeger, 1976, in press

Gregory, P. R , Campbell, J M , and Cheng, B "A Cost-inclusive Simultaneous Equation Model of Birth Rates." *Econometrica* 40 (July 1972) 41–47

Gregory, P. R., and Griffin, J "Secular and Cross-Section Industrialization Patterns. Some Further Evidence on the Kuznets-Chenery Controversy" *Rev Econ. and Statis* 56 (August 1974) 360–68

Griliches, Z "Comment on Nerlove's Paper" *J P E* 82, no 2, suppl. (March/April 1974). S219–S221

Gronau, R. "The Effect of Children on the Housewife's Value of Time." *J P.E.* 81, no 2, suppl (March/April 1973) S168–S199

Hansen, F "Economic Consequences of Rapid Population Growth—a Comment" *Econ J* 50 (March 1972) 110–13

Kelley, A. C "The Role of Population in Models of Economic Growth" *A E.R.* 64 (May 1974) 39–44

Kim, S., et al *Dual Careers A Longitudinal Study of Labor Market Experiences of Women* New York Columbia Univ Press, 1972

Leibenstein, H "The Economic Theory of Fertility Decline" Discussion Paper, Harvard Inst Econ Res , 1973

——— "An Interpretation of the Economic Theory of Fertility Promising Path or Blind Alley?" *J Econ Literature* 12 (June 1974) 457–87

Mincer, J "Market Prices, Opportunity Costs and Income Effects" In *Measurement in Economics*, edited by Carl Christ et al Stanford, Calif Stanford Univ. Press, 1963.

Nelson, R "The Low-Level Equilibrium Population Trap" *A E R* 43 (December 1956) 4364–73

Schultz, T P "An Economic Model of Family Planning and Fertility" *J P.E* 77, no 2 (March/April 1969) 153–80

Simon, J. *The Effects of Income on Fertility* Chapel Hill Univ North Carolina Press, 1974.

Venieris, Y , Sebold, F , and Harper, R "The Impact of Economic, Technological, and Demographic Factors on Aggregate Births" *Rev Econ and Statis* 55 (November 1973) 294–304

Wachter, M L "A Time Series Fertility Equation The Potential for a Baby Boom in the 1970s" Discussion Paper, Univ Pennsylvania, Dept Econ , 1974.

Wickens, M R. "A Note on the Use of Proxy Variables" *Econometrica* 40 (July 1972). 63–69

Willis, R. "A New Approach to the Economic Theory of Fertility Behavior." *J.P.E* 81, no 2, suppl. (March/April 1973) S14–S64.

Testing the Monetary Approach to Devaluation in Developing Countries

Michael Connolly and Dean Taylor

University of Florida, Gainesville

The monetary approach to the balance of payments suggests that surpluses and deficits are essentially adjustment phenomena which serve to equilibrate the money stock in existence to the quantity demanded in an open economy. An increase in the domestic money supply via increased domestic credit creation, by either central or private banks, leads to larger deficits which, in turn, deplete the excess supply of money A devaluation, however, raises domestic prices and as a result increases the demand for money, thereby leading to a temporary improvement in the balance of payments [1]

The purpose of this paper is to test the monetary approach to devaluation from a sample of 18 independent devaluations that took place in developing countries during the 1959–70 period and, at the same time, to indicate the conditions under which devaluation is successful in improving the balance of payments In the process, the paper sheds light on two important related issues first, the degree to which devaluation is inflationary per se, and second, the extent to which devaluation changes the price of traded goods relative to nontraded ones

Part I sets out a simple monetary model appropriate for empirical testing of the monetary approach, Part II summarizes the results of the empirical tests, and Part III focuses upon the impact of devaluation upon the terms of trade and domestic prices The concluding section discusses Richard Cooper's (1971) test of the monetary approach and attempts to assess the weaknesses and strengths of the monetary framework.

I. A Simple Monetary Model

The central proposition of the monetary approach is that *a devaluation will improve the balance of payments unless accompanied by an approximately equiproportionate expansion in domestic credit* This proposition may be sketched

---

We are greatly indebted to Harry G Johnson for encouragement and suggestions at various stages of our work and to Robert A Mundell and an anonymous referee for remarks leading to improvement in the final version of this paper

[1] For a full exposition of the monetary approach, see Johnson (1972)

[*Journal of Political Economy*, 1976, vol 84, no. 4, pt 1]
© 1976 by The University of Chicago. All rights reserved.

more formally. First, assume a stable demand for money function which depends upon domestic prices, p, and the level of permanent income, y:

$$L = kpy, \qquad \text{the demand for money,} \qquad (1)$$

where k, a constant, is the fraction of yearly income that individuals desire to hold in the form of cash balances

Our second assumption is that domestic prices are kept in line with foreign prices via the exchange rate, so that we may write

$$p = rp^*, \qquad \text{purchasing power parity,} \qquad (2)$$

where r is the exchange rate, and p^* the level of world prices

The third relationship we need to specify is the money supply process If we consolidate central and private banks into a single banking system, the total liabilities of the consolidated banking system would equal, by definition, the stock of money, M, plus its nonmonetary liabilities, W The assets of the banking system, on the other hand, can be broken down into two kinds, foreign and domestic, or $R + A$ By deducting nonmonetary liabilities from domestic assets, and defining the difference as domestic credit, or $D = A - W$, the following identity holds [2]

$$M = R + D, \qquad \text{money supply process,} \qquad (3)$$

where M is money narrowly defined, or currency plus demand deposits, R is net holdings of foreign reserves by both central and private banks, and D is domestic credit, or net holdings of domestic assets by the consolidated banking sector Domestic credit can be expanded by the acquisition of additional holdings of domestic assets, either on the part of the central bank or by additional lending by private banks

We assume that the economy adjusts sufficiently rapidly via payments imbalances so that for our purposes money supply equals money demand, or,

$$L = M, \qquad \text{monetary equilibrium} \qquad (4)$$

Finally, for notational purposes, denote the value of permanent income measured at world prices as

$$Y = p^*y, \qquad \text{permanent income valued at world prices} \qquad (5)$$

Substituting (2) and (4) into (1), the resulting expression into (3), making use of (5), and noting that the change in reserves is, by definition, the balance of payments, or $B = \Delta R$, we have, after dividing by M (or krp^*y) $B/M = (1 + g)\Delta r/r + g - \Delta D/M$, where g is the rate of growth of permanent income valued at world prices

Assume that devaluation causes no change in the rate of growth of permanent income measured at world prices, or $g_t = g_{t-1} = g$, where the value of a variable in the period after devaluation is denoted by a subscript t and its value in the period before by $t - 1$. With constant growth in permanent income[3] we may write the improvement in the balance of payments as a proportion of the money stock as

---

[2] This derivation follows Mundell's (1971, pp 90–91)

$$\left(\frac{B_t}{M_t} - \frac{B_{t-1}}{M_{t-1}}\right) = (1 + g)\left(\frac{\Delta r_t}{r_t} - \frac{\Delta r_{t-1}}{r_{t-1}}\right) - \left(\frac{\Delta D_t}{M_t} - \frac{\Delta D_{t-1}}{M_{t-1}}\right). \quad (6)$$

With either a *rigidly pegged* exchange rate or a *continuously depreciating* one, the exchange rate term vanishes and equation (6) states that acceleration in the rate of growth of domestic credit causes the balance of payments to worsen in the same proportion

However, if the change in the exchange rate is not preceded by a prior one, equation (6) assumes the following simple form where the subscripts have been omitted

$$\Delta\frac{B}{M} = (1 + g)\frac{\Delta r}{r} - \Delta\frac{\Delta D}{M} \quad (7)$$

In terms of regression analysis, the coefficient for the step up in the rate of domestic credit creation should be -1, while the coefficient for the rate of devaluation should be slightly greater than 1 with an upward trend in world prices and/or secular growth in permanent income If we assume that full adjustment in prices and the money stock takes place in 1 year, equation (7) will apply for changes over 1-year periods, while if adjustment is slower the same equation can be applied to longer periods In the empirical tests to follow, we estimate equation (7) for 1- and 2-year periods, respectively

II. Testing the Monetary Model

This section reports tests of the simple monetary model for a sample of 18 independent devaluations which took place during the 1959–70 period Insofar as possible, we simply adopted Richard Cooper's sample set.[4]

[3] Theoretically, *permanent* income might rise to a higher level as a result of greater economic efficiency being brought about by the relaxation of controls following devaluation However, our tests using real income data for the 18 countries 5 years before and after devaluation failed to detect such a rise. Nor could we find any systematic relationship between changes in growth rates of *measured* real income, the rate of devaluation, and the increase in the rate of growth of domestic credit for 1- and 2-year periods On the average, real income appears to grow at roughly the same rate following devaluation, independently of the rate of devaluation and the change in the rate of growth of domestic credit

[4] Six countries could not be included since they do not report foreign assets, foreign liabilities, and/or money supply figures on a quarterly basis All of these six devaluations took place in the 1950s for which the data are sparse The countries were Greece, Mexico, Morocco, Pakistan, Peru, Spain, and Turkey Others, such as Brazil and Canada, were not included since changes in the exchange rate were too continuous in nature. The

Along with Cooper, we ruled out the bloc devaluations triggered by the November 1967 devaluation of sterling, and those of the franc zone following the August 1969 devaluation of the franc. Roughly speaking, the sample is composed almost exclusively of small, open, developing countries.[5] The countries and the nominal rates of devaluation are listed in table A1.

Since the countries in the sample report foreign assets and liabilities on a quarterly basis for both central and private banks, a measure of net holdings of foreign reserves (R) by the consolidated banking system is available [6] The change in the net foreign reserve position during the 4 quarters prior to the quarter of devaluation provides a measure of the balance of payments in the year before devaluation, while changes in net foreign reserves the 4 quarters after devaluation measures the payments balance the year after [7] Each measure is then divided by the money stock at the beginning of the respective period The *improvement* in the balance of payments as a proportion of the money stock is then found by simply subtracting the second measure from the first, thereby giving the left-hand side of equation (7) Clearly, the same calculations can be done for 8 quarters before and after devaluation for 2-year comparisons

---

Korean devaluations of February 1960 and February 1961 were too close together for the tests we perform and were thus omitted Iceland's small devaluation of August 1961 was not included for the same reason, but the larger one in February of the preceding year was We then added three devaluations during the decade of the 1960s not included in Cooper's sample—Finland, Peru, and Venezuela—and finally three others that occurred in 1970—Ecuador, Turkey, and Philippines—too late for data to be available for Cooper's study

[5] A major problem with this sort of sample set is that the measure for the *nominal* or official rate of devaluation is typically substantially larger than the *effective* rate as a result of the use of devaluation as a means to eliminate or ease multiple exchange rates, quantitative controls on imports, and restrictions on financial transfers, and to do away with other exchange control schemes such as advance deposits for imports and special foreign exchange accounts for exporters As a result, the effective rate of devaluation invariably falls well short of the official rate Nevertheless, no attempt is made to estimate effective rates of devaluation, primarily for want of a workable way of undertaking the estimates This leads to two sorts of biases that are taken up in detail following the empirical tests In what follows, the rate of devaluation is measured as the percentage increase in the local currency price of the dollar

[6] With few exceptions, the net foreign asset position of the consolidated monetary sector is reported in the 1973 supplement to the *International Financial Statistics* (*IFS*), as are the data on money stocks and exchange rates upon which our calculations are based (for Argentina and Iceland data were taken from the 1965/66 and 1966/67 supplements, respectively).

[7] Since foreign reserves are measured in home currency, the quarter of devaluation is not included in the calculations of the change in reserves so as not to record the accounting revaluation of foreign reserves as part of the improvement in the balance of the payments following devaluation Dropping the quarter of devaluation from the calculations has the additional benefit of diminishing the effect of speculative currency flows in anticipation of devaluation and repatriation afterward for purposes of profit taking. That is, the changes in reserves before and after do not include those changes that take place during the quarter of devaluation.

The definition of money used is the same as the IMF definition, namely, currency plus demand deposits or money narrowly defined. Finally, to measure domestic credit, we deduct net foreign assets of the consolidated banking system, R, from the money stock, M, thereby giving domestic credit, D, from the balance-sheet identity (3).[8]

Regressing the improvement in the balance of payments as a proportion of the money stock on the nominal rate of devaluation and the increase in the rate of growth of credit as a proportion of the money stock, for 1- and 2-year periods, respectively, gives

$$\Delta \frac{B}{M} = \underset{(.06)}{.12} \frac{\Delta r}{r} - \underset{(.14)}{.76} \Delta \frac{\Delta D}{M} \qquad \begin{array}{l} R^2 = .61 \\ \mathrm{SE} = .13, \end{array} \tag{8}$$

$$\Delta \frac{B}{M} = \underset{(.10)}{.29} \frac{\Delta r}{r} - \underset{(.15)}{.82} \Delta \frac{\Delta D}{M} \qquad \begin{array}{l} R^2 = .65 \\ \mathrm{SE} = .20, \end{array} \tag{9}$$

which corresponds to the final expression (7) of the simple monetary model of devaluation. The standard errors are reported in parentheses.

The most striking result common to both regressions is the coefficient of nearly -1 ($-.76$ and $-.82$, respectively) for increases in the rate of growth of domestic credit over 1 and 2 years. The corresponding t-statistics for the hypothesis that the coefficient indeed equals -1 are 1.7 and 1.2, so that one could not reject the hypothesis. Looking at the same result somewhat differently, one would reject the very strong nonmonetary hypothesis that changes in growth in domestic credit play no importance whatsoever since the relevant t-statistics for this test are both -5.4. In short, the evidence tends to support the domestic credit aspect of the monetary model.

However, in both 1- and 2-year periods, the coefficient for the rate of devaluation does not approach anything near unity, and in fact this hypothesis would be rejected. On the other hand, the corresponding t-statistics of 2.1 and 2.9 for 1 and 2 years, respectively, would lead to rejecting the hypothesis that devaluation does not have any effect whatsoever in improving the balance of payments.[9] In short, the evidence on the rate of devaluation is somewhat mixed. However, as noted previously, this is not entirely surprising since devaluation is invariably accompanied

[8] The IMF definition of domestic credit corresponds to ours with nonmonetary liabilities deducted, namely, quasi money and other items net.

[9] We performed the same tests for money more broadly defined so as to include time deposits, or quasi money of the IMF definitions. The coefficients for the rate of devaluation were of roughly the same magnitude as those for currency plus demand deposits reported above. However, the coefficient for the rate of domestic credit creation was $-.48$ for 1 year and $-.18$ for 2 years, the first being significantly different from zero while the second is not. Thus, the monetary model performs well with money narrowly defined but not with money broadly defined. It would appear that the demand for money narrowly defined is more stable than the demand for money broadly defined.

by a substantial relaxation of the paraphernalia of exchange restrictions, which would tend to bias the coefficient for the exchange rate downward. A further bias toward zero results from the errors in variables problem,[10] since one would expect that the magnitude of the effective rate of devaluation relative to the nominal rate varies from country to country. In addition to these biases, there is evidence to be reported in Part III that the major impact of devaluation, at least in the first year, falls upon the prices of traded goods, and only partially upon the general price level. As a result, the demand for money would not be expected to rise by the full proportion of the devaluation, and consequently the coefficient for the rate of devaluation would be lower, particularly for short periods. The larger coefficient for devaluation for the 2-year period would thus reflect, in part, more complete price adjustment.[11]

Our results can be fruitfully depicted in figure 1, where the improvement in the balance of payments as a percentage of the money supply is plotted along the vertical axis and the increase in domestic credit as a proportion of the money stock along the horizontal axis. Graphically, either a continuously depreciating currency or a rigidly pegged one would theoretically yield observations along the line through the origin with a slope of -1. To the extent that a country experiences a discrete devaluation, the observation should lie above the line by a percentage amount slightly greater than the rate of devaluation, namely, $(1 + g)\Delta r/r$ above. Notice that countries that experience improvements in their balance of payments yield points above the horizontal axis, while those with worsened payments positions after devaluation lie below. Similarly, countries expanding the rate of growth in domestic credit yield observations to the right of the vertical axis, and those contracting lie to the left.

---

[10] If the effective devaluation were the same proportion, say 40 percent of the nominal rate, for each and every country, we would expect the estimated coefficient for the rate of devaluation to be approximately .4. However, assume the true relationship between the nominal and effective rate of devaluation is $n = \alpha s + \mu$, where μ is an error term with expectation zero. The value of α in the above example is 2.5, and there is some random component μ causing deviations from the true relationship. This must bias the coefficient for the rate of devaluation toward zero and may bias the coefficient for the rate of growth in domestic credit in either direction (see Levi [1973] on the errors in variables problem for a recent proof). Using a technique described in Theil (1971, pp. 609–10, 614), and reestimating the regressions with a relatively large standard deviation for the error term, $\sigma_\mu = .2$, the corrected estimates of the coefficients for the nominal rate of devaluation and the increase in growth of domestic credit were, for 1 year, +.14 and −.75 and for 2 years, +.34 and −.78, respectively. Since the average nominal devaluation was in fact .48, a standard deviation of the *error term* of .2 would be rather large. With $\sigma_\mu = .1$, the coefficients are nearly identical with those reported in equations (8) and (9). We are indebted to G. S. Maddala for helpful discussion on this point.

[11] The above results can, however, be viewed in a different way. Suppose in fact the monetary model is the correct one and full adjustment takes place over 2 years, equation (9) would imply an average effective rate of devaluation equal to approximately 30 percent of the nominal one.

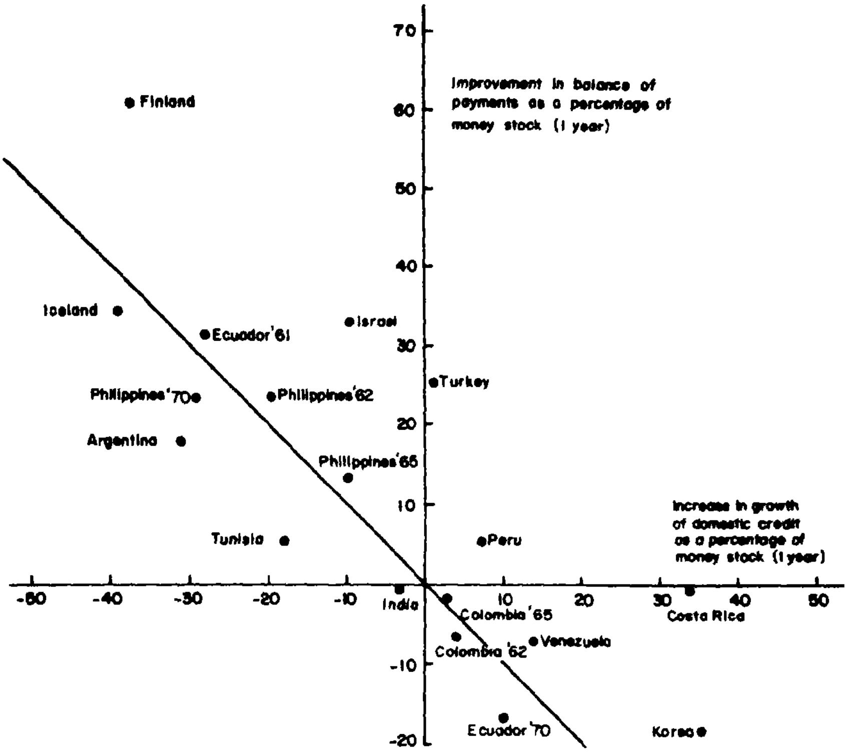


FIG 1

It is clear that the observations do not deviate from the line through the origin by a magnitude near the nominal rates of devaluation This may be the result of a number of things, including the removal of exchange controls And, one would not expect the full impact of the devaluation to take place entirely during the year after This in fact seems to be the case from inspection of figure 2, which illustrates the same relationships for 2 years before and after devaluation In general, the points appear to have drifted upward, indicating more complete adjustment over the two year period

Both figures share one feature in common almost every country which contracted the rate of growth of domestic credit experienced an improvement in the balance of payments over both 1- and 2-year periods That is, the southwest quadrant in both figures is nearly empty It would appear from this that a reduction in the rate of growth in domestic credit is, for all practical purposes, sufficient to guarantee an improvement in the balance of payments.

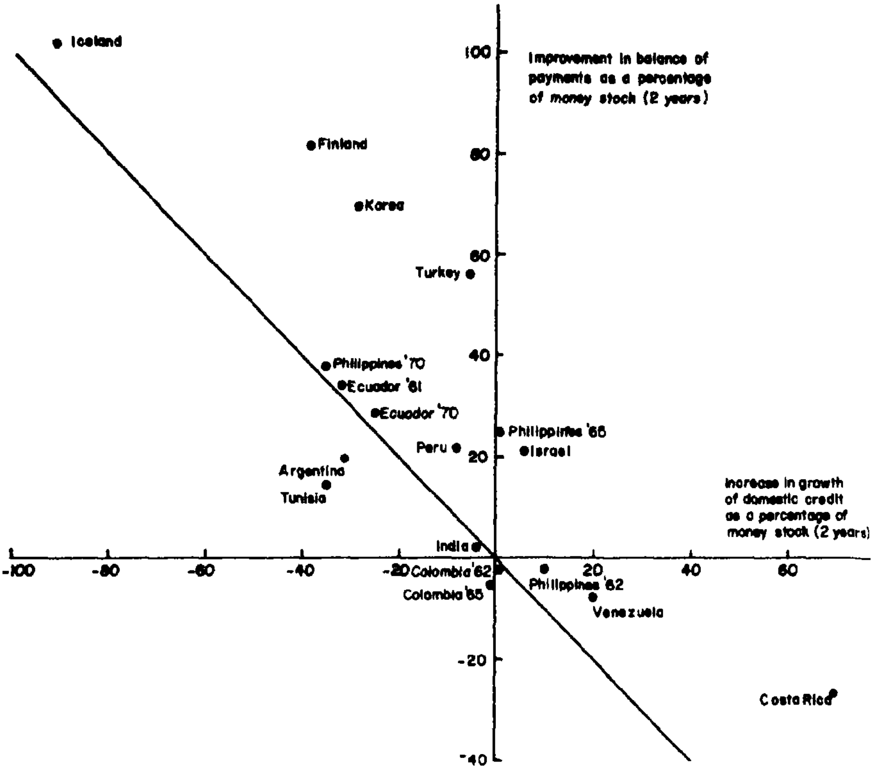


FIG. 2

III. Devaluation and Prices

We have suggested that the price adjustment following devaluation involves primarily a substantial increase in the price of traded goods, exports, and imports and, secondarily, a moderate acceleration in the overall rate of inflation. As a result the coefficient applying to the rate of devaluation should be lower than that suggested by the simple monetary model.

Evidence supporting these assertions is provided by the average behavior of prices before and after devaluation Figure 3 illustrates the path of four price indices—wholesale, consumer, export, and import—in home currency on a quarterly basis for 2 years before and after quarter 0, the quarter of devaluation The indices are calculated for only eight of the devaluations in the sample set since only eight reported *all* four indices, and therefore the results are tentative in nature.[12]

---

[12] The price indices are for Colombia 1962 and 1965; Finland, India, the Philippines 1962, 1965, 1970; and Turkey. All indices are in home currency and are taken from the

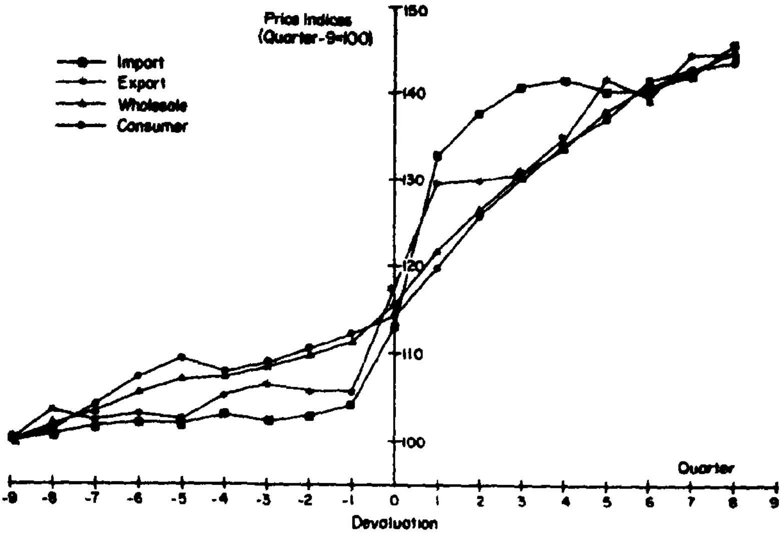


Fig 3

This evidence exhibits the following characteristics:

1. Export and import prices rise very slightly prior to devaluation at about the same rate, but wholesale and consumer prices rise at a faster but nearly identical rate

2 In the year following devaluation, the prices of exports and imports rise substantially, while consumer and wholesale prices rise at a faster rate than before but at a slower rate than the prices of traded goods

3 During the second year after devaluation, *all* prices rise at about the same rate

4 Over the *entire* 2-year period after devaluation, the prices of exports and imports rise at a greater rate than wholesale and consumer prices, and this increase in the relative price of traded goods was sufficient to rectify the relative decline in their price prior to devaluation

If we suppose that the rates of inflation of each index would have remained the same had no devaluation taken place, then the devaluation would have added 13 and 14½ percent to the consumer and wholesale price indices, and 30 and 31½ percent to the export and import price indices over 2 years. For the eight observations, the average rate of devaluation was 51 percent, so that while the prices of traded goods may have risen by an amount near the devaluation, prices in general did not.

1973 supplement to the *IFS*, except those for export and import prices for India, which are from the December 1967 and December 1968 issues. Each price index is calculated by averaging across countries for each quarter the level of the corresponding price index.

In summary, it would appear that devaluation in developing countries does not have substantially inflationary effects The view that inflation brings about devaluation might appear more credible than the reverse.

IV. Concluding Remarks

In assessing the monetary framework we find·

1 Devaluation in developing countries appears to be quite successful in improving payments imbalances, particularly in light of the fact that it frequently serves the second objective of dismantling exchange controls Further, the extent to which devaluation is successful depends in large part upon the monetary policy pursued along with the change in the exchange rate In particular, a moderate decline in the rate of growth of domestic credit seems sufficient to insure that devaluation improves the balance of payments

2 On the other hand, the price adjustment following devaluation involves primarily a substantial increase in the price of traded goods and, secondarily, a moderate increase in the overall rate of inflation In most instances, devaluation is preceded by inflation and in some by falling export prices, so that the change in the exchange rate is perhaps best viewed as a policy response to monetary and/or real disturbances depending upon the particular instance In light of the related fact that exchange controls are frequently relaxed following devaluation, it is not surprising that its consequences are only moderately inflationary relative to rates of devaluation However, the improvement in the balance of payments and the rate of devaluation are positively related in both 1- and 2-year tests The coefficients for the rate of devaluation, while significant, fall well short of unity, however, so that the price-adjustment mechanism of the simple monetary model would be rejected

3 In conclusion, we feel it important to relate our generally positive findings on the monetary approach to the negative ones reported by Richard Cooper (1971) in his extremely valuable study. Specifically, Cooper found no correlation between the change in domestic credit of the banking system in the year following devaluation and the improvement in the goods and service balance in the same year for roughly the same sample of countries we analyze His finding differs from ours primarily in that his is for only the merchandise account, while the monetary approach has to do with the overall balance of payments The monetary argument is basically that one cannot expect devaluation to improve the overall balance of payments unless it leads to an increased demand for money unmatched by an increase in domestic credit. His test suffers from two other shortcomings: first, the magnitude of the rate of devaluation by country is not taken into account, and second, as pointed out by Carlos F. Diaz Alejandro in his comments on Professor Cooper's paper

(Cooper 1971, p 513), the comparison should involve *increases* in the rate of growth of domestic credit following devaluation, as is done in our study, rather than the *rate* of growth of domestic credit, as in his study.

Appendix

TABLE A1

DEVALUING COUNTRIES

	Country	Date	Nominal Devaluation in Home Currency
1	Argentina	January 1959	58
2	Colombia	November 1962	34
3	Colombia	September 1965	50
4	Costa Rica	September 1961	18
5	Ecuador	July 1961	20
6	Ecuador	August 1970	39
7	Finland	October 1967	30
8	Iceland	February 1960	56
9	India	June 1966	58
10	Israel	February 1962	67
11	Korea	May 1964	96
12	Peru	October 1967	44
13	Philippines	January 1962	94
14	Philippines	November 1965	11
15	Philippines	February 1970	65
16	Tunisia	September 1964	24
17	Turkey	August 1970	65
18	Venezuela	January 1964	42

SOURCE —Unless otherwise noted, the rates of devaluation are taken from the 1973 supplement to the *International Financial Statistics* Argentina (1) *from free rate in third quarter* of 1958 to official rate in second quarter of 1959, 1965/66 supplement to *IFS*, p 4, Iceland (8) from midpoint of selling range in quarter prior to devaluation to new selling rate 1966/67 supplement to *IFS* p 120, Philippines (14) from *IFS* (December 1966), p 245

References

Cooper, Richard N "An Assessment of Currency Devaluation in Developing Countries " In *Government and Economic Development*, edited by Gustav Ranis New Haven, Conn Yale Univ Press, 1971

International Monetary Fund *International Financial Statistics* Washington IMF, various issues and supplements

Johnson, Harry G "The Monetary Approach to Balance of Payments Theory " *J Financial and Quantitative Analysis* 7, no 2 (March 1972) 1555–72. Also in *International Trade and Money*, edited by Michael Connolly and Alexander Swoboda Toronto Univ Toronto Press, 1973

Levi, Maurice D "Errors in the Variables Bias in the Presence of Correctly Measured Variables " *Econometrica* 41, no 5 (September 1973) 985–86.

Mundell, Robert A *Monetary Theory* Pacific Palisades, Calif Goodyear, 1971

Theil, Henri *Principles of Econometrics* New York Wiley, 1971

Another Advantage of the Division of Labor

Nathan Rosenberg

Stanford University

It is almost unseemly still to be searching out further advantages which Adam Smith attributed to the division of labor Surely the 200 years which have now elapsed since the publication of the *Wealth of Nations* has been ample time for even the dullest-witted assortment of scholars to ferret out all interesting bits of Adam Smith's discussion of that subject. Nevertheless the purpose of this paper will be to call attention to yet another significant advantage beyond the three which Adam Smith identified in the celebrated first chapter of the *Wealth of Nations*[1]

The discussion with which I will be concerned did not appear in the *Wealth of Nations* but in a set of lectures on rhetoric and belles lettres which Adam Smith delivered at the University of Glasgow in 1762–63 (Smith 1971) (Smith scholars who might experience some self-mortification at their failure to perceive the significance of this treatment may therefore be reassured that the lectures have been in the public domain for only a few years) I intend to show that this discussion is not only neatly consistent with and supplementary to the time-honored treatment in the *Wealth of Nations*, but that it also enlarges our understanding of Smith's views in some important respects

The material with which I am concerned was presented in a lecture on February 7, 1763[2] In earlier lectures Smith had been concerned with

---

The author wishes to acknowledge the valuable comments which he received on an earlier draft from Paul David, David Mowery, and George Stigler

[1] "Thus great increase of the quantity of work, which, in consequence of the division of labour, the same number of people are capable of performing, is owing to three different circumstances, first, to the increase of dexterity in every particular workman, secondly, to the saving of the time which is commonly lost in passing from one species of work to another, and lastly, to the invention of a great number of machines which facilitate and abridge labour, and enable one man to do the work of many" (Smith 1937, p 7) For a further discussion of Smith's views on the division of labor, see Rosenberg (1965)

[2] Interestingly enough, these lectures were delivered at about the same time as his *Lectures on Justice, Police, Revenue and Arms* (Smith 1956) On the exact dating of those student lecture notes Cannan concluded "It is probable that the actual lectures from which the notes were taken were delivered either in the portion of the academical session of 1763–4 which preceded Adam Smith's departure, or in the session of 1762–3, almost certain that they were not delivered before 1761–2, and absolutely certain that they were not delivered before 1760–1" (Smith 1956, p. xx)

[*Journal of Political Economy*, 1976, vol 84, no 4, pt 1]
© 1976 by The University of Chicago. All rights reserved

demonstrative and deliberative eloquence and in this lecture turned his attention to judicial eloquence. Legal arguments, he asserts, take essentially two forms, either to demonstrate by "abstract reasoning" that something "follows from some statute," or by showing "how it has been supported as law by former practice and similar adjudged causes or precedents" (1971, p 168) At this juncture Smith points out that neither the Greek nor Roman lawyers had made any use of the second category, which is so widely exploited by modern lawyers, and he undertakes in a digression to explain this failure His explanation turns upon a more extensive division of labor in modern as compared with ancient times

A central feature of earlier periods, Smith points out, is that the duties of judge, general, and legislator were commonly performed by the same persons—"at least the two former are very commonly conjoined" (1971, p 168). In barbarous societies, when men formed the habit of submitting themselves to control by some superior person, they did not generally distinguish between different realms—military, legislative, or judicial As a result,

> The same persons therefore who judged them in peace, led them also to battle In this two-fold capacity of judge and general, the first kings and consuls of Rome, and other magistrates, would reckon the judicial part of their office a burthen, rather than that by which they were to obtain honour and glory, that was only to be got by military exploits They therefore were very bold in passing sentence They would pay very little regard to the conduct of their predecessors, as this was the least important part of their office This part was therefore for their ease separated from the other and given to another sort of magistrates These, as the Judicial was their only office, would be at much greater pains to gain honour and reputation by it [Smith 1971, p 168, cf Smith 1937, pp 680–81][3]

In this passage Smith thus argues in favor of a more extensive division of labor by asserting that it is better to have two important social functions divided up among separate individuals rather than having them combined in a single individual. The reasoning underlying the argument will be familiar to all readers of the *Wealth of Nations*. If an individual's attainment of recognition and success is linked to only one of the two activities which he regularly performs, he will have no particular motivation to perform the other activity well A man who is simultaneously general and judge should not be expected to devote any extensive effort to the latter function if his place in society is determined entirely by his qualities as a military leader It hardly requires any marginalist calculations to predict that, if the returns to activity A remain very high while those for activity B are effectively set at zero, the latter activity will be either performed in a most perfunctory way or totally neglected One can almost hear the Adam Smith of the *Wealth of Nations* intoning at this point. "In every profession, the exertion of the greater part of those who exercise it, is always in proportion to the necessity they are under of making that exertion" (1937, p 717) If there is no necessity or no gain to be derived from the performance of the judicial function, it cannot reasonably be expected that it will be performed well[4]

---

[3] Of course Smith also recognized that the exercise of the judicial power on the part of the sovereign was likely to lead to many grave abuses, especially when the administration of justice was made "subservient to the purposes of revenue " See Smith (1937), bk 5, chap 1, pt 2, "Of the Expense of Justice," where he observes "In all barbarous governments, accordingly, in all those ancient governments of Europe in particular, which were founded upon the ruins of the Roman empire, the administration of justice appears for a long time to have been extremely corrupt, far from being quite equal and impartial even under the best monarchs, and altogether profligate under the worst" (p. 676) More generally "When the judicial is united to the executive power, it is scarcely possible that justice should not frequently be sacrificed to, what is vulgarly called, politics. . In order to make every individual feel himself perfectly secure in the possession of every right which belongs to him, it is not only necessary that the judicial should be separated from the executive power, but that it should be rendered as much as possible independent of that power" (p 681)

From this point of view, therefore, a new division of labor involving the separation of military and judicial functions makes great sense since those who now perform only the judicial function have a powerful incentive to perform it well Under the old division of labor, by contrast, judicial responsibilities were in the hands of people who lacked the motivation to perform them well Indeed, this new division of labor is freighted with the greatest of all social consequences The establishment of a separate judiciary not only creates new and powerful motives for the efficient performance of judicial functions It also institutionalizes forces which lead to a highly desirable pattern of cumulative social change For a full-time judiciary

> would be at pains to strengthen their conduct by the authority of their predecessors When, therefore, there were a few judges appointed, these would be at great pains to vindicate and

---

[4] As Smith states elsewhere about a university teacher who derives his entire income from endowments and is therefore totally independent of any form of remuneration from his pupils "His interest is, in this case, set as directly in opposition to his duty as it is possible to set it. It is the interest of every man to live as much at his ease as he can, and if his emoluments are to be precisely the same, whether he does, or does not perform some very laborious duty, it is certainly his interest, at least as interest is vulgarly understood, either to neglect it altogether, or, if he is subject to some authority which will not suffer him to do this, to perform it in as careless and slovenly a manner as that authority will permit If he is naturally active and a lover of labour, it is his interest to employ that activity in any way, from which he can derive some advantage, rather than in the performance of his duty, from which he can derive none" (1937, p 718)

> support their conduct by all possible means. Whatever, therefore, had been practised by other judges would obtain authority with them, and be received in time as law. This is the case in England. The sentences of former cases are greatly regarded, and form what is called the Common Law, which is found to be much more equitable than that which is founded on Statute only, for the same reason as what is founded on practice and experience must be better adapted to particular cases than that which is derived from theory only. [Smith 1971, pp. 168–69]

Once the division of labor has led to the establishment of an independent judiciary, other dynamic forces are set into motion. For, as Smith emphasizes in the *Wealth of Nations*, the pursuit of private self-interest on the part of judges now generates a competitive process the outcome of which is a judicial system far more responsive to individual needs:

> The fees of court seem originally to have been the principal support of the different courts of justice in England. Each court endeavoured to draw to itself as much business as it could, and was, upon that account, willing to take cognizance of many suits which were not originally intended to fall under its jurisdiction. . . . each court endeavoured, by superior dispatch and impartiality, to draw to itself as many causes as it could. The present admirable constitution of the courts of justice in England was, perhaps, originally in a great measure, formed by this emulation, which anciently took place between their respective judges, each judge endeavoring to give, in his own court, the speediest and most effectual remedy, which the law would admit, for every sort of injustice. [1937, p. 679, see also 1956, p. 49]

Note that, in this particular case at least, the competitive process (the endeavor of individuals to widen their jurisdiction) works *against* greater specialization. But justice is, in fact, a very peculiar kind of service, and its dispensation cannot be safely left to the marketplace. The freedom of individuals to "purchase" justice may lead to the destruction of justice itself, especially in the all-too-likely circumstances where the sovereign (or judge) treated judicial service as a source of revenue.[5]

---

[5] "This scheme of making the administration of justice subservient to the purposes of revenue, could scarce fail to be productive of several very gross abuses. The person, who applied for justice with a large present in his hand, was likely to get something more than justice; while he, who applied for it with a small one, was likely to get something less. Justice too might frequently be delayed, in order that this present might be repeated. The amercement, besides, of the person complained of, might frequently suggest a very strong reason for finding him in the wrong, even when he had not really been so. That such abuses were far from being uncommon, the ancient history of every country in Europe bears witness" (Smith 1937, p. 675).

For the free operation of market forces may make it possible for affluent litigants to buy favorable decisions from supposedly impartial judges. The problem is that the enforcement of justice requires that its dispensation be linked institutionally to the power of the state,[6] whereas such linkage inevitably creates numerous opportunities and inducements for the corruption of the service itself It is an extremely delicate matter to organize the administration of justice in such a way that outcomes are not influenced by the desire for increased revenue on the part of either the judge or the sovereign.[7]

There is another aspect of the administration of justice to which Smith also calls attention Even when the military and judicial functions have been effectively separated, it is further essential that the judicial function not be placed in the hands of large social units. For it is unlikely that an activity will be performed well if individual motivations are not strongly engaged to perform it well And when individuals share responsibilities within a sufficiently large collectivity, individual incentives are inevitably diffused or even totally dissipated [8] In judicial proceedings, as elsewhere, it is in society's interests to maintain the closest possible linkage between individual effort and individual reward To do this properly therefore requires that the actions of single individuals be highly visible·

> These judges, when few in number, will be much more anxious to proceed according to equity than where there is a great number The blame then is not so easily laid upon any particular person, they are in very little fear of censure, and are out of danger of suffering much by wrong proceedings Besides that a great number of judges naturally confirm each other, prejudice and inflame each other's passions We see accordingly that the sentences of the judges in England are greatly more equitable than those of the Parliament of Paris or other Courts which are severed from censure by their number The House of Commons, when they acted in a Judicial capacity, have not always proceeded with the greatest wisdom, although their proceedings are kept upon record as well as those of the other Courts, and without doubt in imitation of them [Smith 1971, p 169][9]

---

[6] Although there is a form of competition, just described, among existing courts, there is no free entry into the justice trade, which is monopolized by the state

[7] It is this problem which accounts for Smith's central concern in "Of the Expense of Justice" (n 3 above).

[8] Is this an eighteenth-century anticipation of the "free rider" problem?

[9] Similar difficulties had plagued the ancient world "The case was the same with regard to the Areopagus and the Council of Five Hundred at Athens Their number was too great to restrict them from arbitrary and summary proceedings They would here pay as little regard to the proceedings of former judges as those did who at the same time possessed the office of general along with that of judge The praetor at Rome, indeed, often borrowed from the decrees; but then nothing could be quoted as law to him but what

In his judicial digression Smith therefore emphasizes two different kinds of advantages in connection with an increased division of labor. When responsibilities are divided up among different people so that no one possesses both military and judicial obligations, both obligations will be better performed. Judicial responsibilities are performed better because they now constitute someone's sole responsibility, and that person is being judged solely by how well he performs it. When an individual performs more than one function, but his success is based only upon his performance in one of them, he will neglect the one to which his social status is less attached On the other hand, it is essential that important social functions not be reposed in large social aggregates. A division of labor involving fewness is essential because fewness is a precondition for establishing individual responsibility Only when individual responsibility (or fault) can be established will individual effort be reliably engaged in ways which will promote society's larger interests[10] Thus, whereas Adam Smith had described the advantages of the division of labor in the opening chapter of the *Wealth of Nations* in terms of an improvement in human skills and capacities, his treatment in his lectures on rhetoric and belles lettres calls attention to basic improvements in motivation as well

When law becomes a highly specialized activity, then, subject to an extensive division of labor and administered by small decision-making units, individual motivations and capacities are more deeply and effectively involved in establishing justice and providing the inhabitants of a country with liberty and security The point is not of secondary or trivial interest to Smith Indeed, it is tempting to say that no point was *more* interesting to him, because justice is the essential and indispensable cement which binds a society together Beneficence, as he was at great pains to establish elsewhere, is not enough "Beneficence is less essential to the existence of society than justice Society may subsist, though not in the most comfortable state, without beneficence, but the prevalence of injustice must utterly destroy it" (1880, p. 79).

In this vital way, then, is the history of the increasing division of labor

---

was found in his edict, which was put up at the beginning of each year, and in which he declared in what manner he was to regulate his conduct. (This was the custom till the time of the Edictum Perpetuum) He would have taken it as a great affront to his judgement to have been told that such an one before had done so-and-so And no part of the former edicts could be quoted but what was transcribed into his, and in his name it was always to be quoted There was therefore no room for precedents in any Judicial pleadings amongst the Greeks or Romans . " (Smith 1971, pp. 169–70)

[10] It is important to recognize here that Smith is dealing with two separate issues the extent of the division of labor and the size of the appropriate decision-making unit. In his *Lectures on Rhetoric and Belles Lettres*, however, Smith joins these issues by insisting that fewness is necessary to derive the benefits which might be expected to flow from greater division of labor in judicial proceedings. Note that on p. 169 of those lectures (quoted above) he had stated that judges would be "anxious to proceed according to equity" when they were "few in number." This important qualification, as we will see below, is repeated in the next paragraph (1971, p. 170).

in the judicial process inseparably linked to the fulfillment of other goals which constitute the foundations of any truly civilized community. For beneficence, however desirable and meritorious, is never a sufficiently strong force:

> Though Nature... exhorts mankind to acts of beneficence, by the pleasing consciousness of deserved reward, she has not thought it necessary to guard and enforce the practice of it by the terrors of merited punishment in case it should be neglected. It is the ornament which embellishes, not the foundation which supports, the building, and which it was, therefore, sufficient to recommend, but by no means necessary to impose Justice, on the contrary, is the main pillar that upholds the whole edifice. If it is removed, the great, the immense fabric of human society, that fabric which to raise and support seems in this world, if I may say so, to have been the peculiar and darling care of Nature, must in a moment crumble into atoms [Smith 1880, p 79]

Finally, by joining the hitherto separate threads of Smith's treatment of the division of labor and the administration of justice, we arrive at a deeper appreciation of what Smith had in mind when he observed in chapter 4 of book 3 of the *Wealth of Nations* that " commerce and manufactures gradually introduced order and good government, and with them, the liberty and security of individuals, among the inhabitants of the country, who had before lived almost in a continual state of war with their neighbors, and of servile dependency upon their superiors. This, though it has been the least observed, is by far the most important of all their effects (1937, p 385)[11] Although it is possible to infer some of Smith's meaning when the statement is placed in the larger context of the argument of the *Wealth of Nations* as a whole,[12] we can now make use of the student's lecture notes to draw the highly significant conclusion to which Smith intended his argument to lead, that "... it may be looked upon as one of the most happy parts of the British Constitution, though introduced merely by chance and to ease the men in power, that the office of judging causes is committed into the hands *of a few persons* whose sole employment it is to determine them (1971, p. 170; my emphasis). For increasing wealth has led, in the British case, via an increasing division of labor, to improvements in the administration of justice of the greatest moment for society. It seems appropriate to leave the last word in this matter to Smith himself:

---

[11] In the next sentence Smith adds "Mr Hume is the only writer who, so far as I know, has hitherto taken notice of it " The relationship is, of course, a reciprocal one, in that commerce and manufactures can continue to flourish only in a society where justice continues to be properly administered "Commerce and manufactures can seldom flourish long in any state which does not enjoy a regular administration of justice, in which the people do not feel themselves secure in the possession of their property, in which the faith of contracts is not supported by law, and in which the authority of the state is not supposed to be regularly employed in enforcing the payment of debts from all those who are able to pay" (1937, p 862)

[12] The development of freedom and justice depended, historically, upon the destruction of the power and authority of the ancient barons of the Middle Ages and the end of the "servile dependency" of their retainers and tenants upon them Smith spells out in the *Wealth of Nations* (bk 3, chap 4) how the rise of commerce destroyed the power of the ancient barons and thereby "gradually introduced order and good government." Along with this went the elimination of the arbitrary exercise of power "The tenants having in this manner become independent, and the retainers being dismissed, the great proprietors were no longer capable of interrupting the regular execution of justice, or of disturbing the peace of the country" (1937, p 390, see also 1956, p 155) "Nothing tends so much to corrupt mankind as dependency, while independency still increases the honesty of the people. The establishment of commerce and manufactures, which brings about this independency, is the best police for preventing crimes" (see also Rosenberg 1968)

> The separation of the province of distributing justice between man and man from that of conducting affairs and leading armies, is the great advantage which modern times have over ancient, and the foundation of that greater security which we now enjoy, both with regard to liberty, property, and life. It was introduced only by chance to ease the supreme magistrate of the most laborious and the least glorious part of his power, and has never taken place until the increase of refinement and the growth of society have multiplied business immensely [1971, p. 170; cf 1937, pp 680–81]

References

Rosenberg, N "Adam Smith on the Division of Labour · Two Views or One?" *Economica* 32 (May 1965). 127–39.

———. "Adam Smith, Consumer Tastes, and Economic Growth." *J.P.E* 76 (May/June 1968). 361–74.

Smith, A *The Theory of Moral Sentiments*, as reprinted in *Essays Philosophical and Literary*. London: Ward, Lock, 1880.

———. *The Wealth of Nations*. New York · Modern Library, 1937.

———. *Lectures on Justice, Police, Revenue and Arms, Delivered in the University of Glasgow, Reported by a Student in 1763*. Edited by Edwin Cannan. New York. Kelley & Millman, 1956

———. *Lectures on Rhetoric and Belles Lettres, Delivered in the University of Glasgow, Reported by a Student in 1762–63* Edited by John M. Lothian 1963. Carbondale. Southern Illinois Univ Press, 1971.

Taxes, Spending, and Property Values: A Comment and Further Results

George Richard Meadows

University of Wisconsin—Milwaukee

In a recent issue of this *Journal*, Matthew Edel and Elliott Sclar (1974) present an interesting criticism and extension of Wallace Oates's well-known article (1969). Edel and Sclar argue that the cross-section relationship between property values and local public budgets reported by Oates does not confirm the Tiebout (1956) hypothesis of the existence of an approximation to an efficient allocation of local public goods and services. Rather, Oates's results merely confirm that families have preferences regarding local service levels and act accordingly. Moreover, Edel and Sclar continue, the presence of a significant negative relationship between residential property values and the local effective property tax rate (property tax capitalization) implies that Tiebout's solution is not achieved —at least not for the period covered by the data. Nevertheless, Edel and Sclar suggest that, by observing changes in the magnitude of property tax capitalization over time in a given housing market, we may be able to discern movement toward the Tiebout result as supplies of local public services are adjusted to their long-run equilibrium values. They conclude by presenting results for the Boston area which seem to suggest movement toward Tiebout equilibrium for local public education between 1950 and 1970.

This paper attempts to demonstrate that although Edel and Sclar have made an important contribution in the theoretical investigation of this issue their empirical methodology suffers from two major shortcomings. First, they have failed to account adequately for several factors resulting in intercommunity differentials in local property taxes and in residential property values, second, the structure of their empirical model, and its interpretation, is inappropriate to the hypothesis being considered. Finally, it is shown that when these deficiencies are corrected in a reexamination of the New Jersey sample utilized by Oates, the results strongly suggest an approximation to the Tiebout result in both 1960 and 1970.

I am grateful for the comments and suggestions of Swarnjit Arora, Douglas Booth, Joseph DeSalvo, Manuel Gottlieb, Tong Lee, Ulaganathan Sankar, and a referee.

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved.

Problems with Testing the Tiebout Hypothesis

The immediate difficulty to be faced when we attempt to apply the Tiebout model to modern urban reality is, as Edel and Sclar put it, "holding other background conditions constant." (Oates, of course, was the first to address this problem empirically [1969, p. 960].) That Edel and Sclar have not done enough in this important area can be appreciated by considering the following omissions in their empirical analysis. (1) they have not allowed for the possibility of a nonuniform intercommunity mix of residential and nonresidential property; (2) intercommunity differentials in state and federal aid to education and in the proportion of school-age children enrolled in the public schools are ignored; (3) they make no attempt to hold constant variations in basic physical characteristics of the housing stock such as the age and size of structures; and (4) in their use of nominal property tax rates, they have failed to adjust for the likely intercommunity differences in assessment/sales ratios. These omissions raise serious questions regarding the reliability of the estimates of tax and expenditure capitalization reported by Edel and Sclar.

Nor do the difficulties end here. Edel and Sclar note that Oates's findings of partial capitalization of both property tax rates and school expenditures do not support his conclusion regarding the efficacy of the Tiebout allocation mechanism; they assert that any capitalization implies a less than full Tiebout equilibrium. Although Edel and Sclar's theoretical assertion is valid, as a generally applicable test of the Tiebout hypothesis their single-equation model can be challenged on two additional grounds. First, they appear to deny the possibility of reverse causality between property values, property tax rates, and local expenditures which Oates (1969, pp. 964–65) argues necessitates the use of a simultaneous-equations model. Second, without a simultaneous-equations model they are unable to isolate and evaluate the Tiebout-relevant indirect impacts on property values caused by a change in any one fiscal variable.

To see this second point, consider the budget policy alternatives confronting a representative municipal governing body in a highly differentiated "Tiebout-type" urban area. Suppose that these individuals actually want to implement a local budget which reflects the preferences of the existing residents. In the annual budget deliberations, each new spending proposal (or proposal that spending be curtailed) will be considered from the vantage point of whether or not it will, on balance, make that community a more desirable place to live.[1] The local governing body will have four major classes of action open to it each year (all in real terms): (1) keep the same budget it had last year, (2) keep the same tax rate but change the mix of local services, (3) keep the same mix of local services but change the tax rate, and (4) change both the tax rate and the mix. If we assume that this decision is reached in any one community on the assumption that other neighboring communities maintain their existing budgets, then the alternative selected may be thought of as that which increases (or prevents the decline in) property values the most. It could be concluded that a Tiebout equilibrium had been reached when the only alternative which prevents a decline in site values is to keep things as they are

---

[1] It should be noted that these decisions must always be made in the context (anticipation) of the local public budget decisions being made in other neighboring communities in the same housing market.

Suppose, however, a given community finds itself out of Tiebout equilibrium and that movement toward the equilibrium requires some change in the local budget Any such budgetary change cannot be confined to only one fiscal variable, therefore, any empirical investigation of the Tiebout hypothesis must at least attempt to encompass the additional (and usually offsetting) effects on property values stemming from the other fiscal changes For example, the decision to reduce local property taxes, *ceteris paribus*, is also a decision to reduce local service levels Likewise, the decision to increase school expenditures is also the decision to reduce nonschool expenditures unless tax rates are also changed. Accepting this and Oates's simultaneity argument, it follows that the full fiscal impact on property values from a change in the local budgetary decision will not be reflected in the coefficient estimates from a single structural equation—not even when that single equation is embedded in a simultaneous-equation model (as in Oates) Generally speaking, the Tiebout hypothesis can only be tested by examining the full effects (both direct and indirect) on property values caused by a change in tax rates and local service levels

Edel and Sclar's single-equation model provides a valid test of the Tiebout hypothesis only when simultaneity can be denied and where intercommunity tax rate and expenditure differentials result solely from differences in tastes Under these circumstances, Tiebout equilibrium, at least for all public goods considered at once, would be indicated by the absence of tax and expenditure capitalization as Edel and Sclar suggest But in the absence of these two conditions, property tax and expenditure capitalization indicated in a single equation may not be inconsistent with Tiebout equilibrium for either all public goods markets considered at once or for individual public services considered separately.

A Simultaneous-Equations Model for Testing the Tiebout Hypothesis

I present a four-equation model in which the first equation is similar to that employed by Oates and other previous investigators. This model differs from earlier ones, however, in that the generally acknowledged simultaneity among property values, property taxes, and local expenditures is given an explicit structural form. The structural equations and variable definitions are as follows:

$$V = a_0 + a_1 T + a_2 E_s + a_3 E_m + a_4 R + a_5 D + a_6 A + \mu_1, \tag{1}$$

$$\begin{aligned} T = {} & b_0 + b_1 E_s + b_2 E_m + b_3\, RVALPR + b_4\, NRVALPR \\ & + b_5\, AIDPP + b_6\, PUB + \mu_2, \end{aligned} \tag{2}$$

$$\begin{aligned} E_s = {} & c_0 + c_1 T + c_2 E_m + c_3\, RVALPP + c_4\, NRVALPP \\ & + c_5\, AIDPP + c_6\, PUB + c_7\, HS + \mu_3, \end{aligned} \tag{3}$$

$$\begin{aligned} E_m = {} & d_0 + d_1 T + d_2 E_s + d_3\, RVALPR + d_4\, NRVALPR \\ & + d_5\, OWN + d_6\, DEN + d_7\, X + d_8\, NWK \\ & + d_9\, AREA + \mu_4, \end{aligned} \tag{4}$$

where:

V = median value of owner-occupied dwellings,
T = equalized property tax rates for all local services including education,
E_s = weighted school expenditures per pupil (operating and debt service),
E_m = municipal expenditures per resident,
R = median number of rooms per residential dwelling unit,
D = index of accessibility (weighted average linear distance to area central cities),
A = average age of residential dwelling units,
$RVALPR$ = equalized value of residential (owner and non-owner occupied) property per resident,
$NRVALPR$ = equalized value of nonresidential property per resident,
$AIDPP$ = state and federal aid to education per pupil,
PUB = percentage of resident school-age children attending public schools,
$RVALPP$ = equalized value of residential property per pupil,
$NRVALPP$ = equalized value of nonresidential property per pupil,
HS = percentage of residents over 25 years old having completed high school,
OWN = percentage of residential dwelling units owner occupied,
DEN = population density,

X = percentage of real property value exempt from property taxation,
NWK = percentage of residents 16 years old or over who are not members of the labor force,
$AREA$ = land area of the municipality,
μ_i = ith equation's stochastic disturbance term

Equation (1), where residential property values are determined by local property tax rates, public expenditure levels, average size of dwelling units, accessibility to regional employment centers, and average age of dwelling units, is quite similar to that employed by Oates and other previous investigators. In equation (2), local general property tax rates are determined by local expenditure levels, local fiscal capacity (residential and nonresidential property), state and federal aid to education, and the ratio of school-age children attending public schools. These more inclusive measures of municipal fiscal capacity are believed to be superior to that used by Oates because the residential-nonresidential mix and the proportion of owner-occupied units varies widely across his New Jersey sample.[2] Moreover, certain nonresidential properties may convey fiscal advantages upon the residents of the community, viz, reductions in the tax price of local services.[3] If this variable (say nonresidential property per resident) is positively correlated with property values and tax rates or local service levels, its omission will bias upward the regression coefficient estimates of the included fiscal variables, indicating a greater degree of capitalization than actually exists.[4] The same argument can be made with respect to the cost of local public education, because these costs will vary inversely with state and federal aid and nonpublic school attendance.[5] In equation (3), school expenditures per pupil are a function of the percentage of resident adults having completed high school and the previously discussed fiscal variables. In equation (4), local expenditures for municipal services per capita are determined by the percentage of owner-occupied dwelling units, population density, the proportion of real property exempt from property taxation, the percentage of working-age residents not in the labor force, the land area of the community, and the fiscal variables.

---

[2] In Oates's sample of 53 communities in 1960, the percentage owner occupied varies from 6.3 in West New York to 91.5 in Glen Rock—$\bar{X}$ = 62.03, SD = 19.73.

[3] Any nonresidential activity will be desirable if its tax payments (or contributions in lieu of taxes) to the local public sector equal or exceed the marginal costs of the local public services it receives and if there is no net decline in the community's attractiveness as a place to live through changes in such factors as accessibility to employment, shopping, and entertainment centers and changes in environmental quality (air, noise, and aesthetic external effects).

[4] See Pollakowski (1973, pp. 995–98) and Edelstein (1974, p. 323) for a discussion of this problem.

[5] Edel and Sclar (1974) mention this possibility in their paper.

The task of calculating the full impact on property values of any local budgetary change is further complicated by the fact that the variables we are primarily interested in are themselves endogenous. If they were exogenous instead, the full simultaneous impact on property values could be directly calculated from the reduced form of a simultaneous-equation model. Although we will not be able to calculate the full effect stemming from a change in any one of our endogenous variables, it will be possible to calculate what may be thought of as a "first approximation" of the full effect. For example, with a change in taxes, the effect on property values is approximated by

$$dV/dT = \partial V/\partial T + (\partial V/\partial E_s)(\partial E_s/\partial T) + (\partial V/\partial E_m)(\partial E_m/\partial T) \qquad (5)$$

Equation (5) incorporates the inevitable changes in school expenditures E_s and municipal expenditures E_m which accompany any change in local taxes and which will themselves also affect property values. All of the partial derivatives in (5) can be estimated directly from the structural equations (1) – (4). Thus rewriting (5) in terms of estimated regression coefficients gives

$$dV/dT = \hat{a}_1 + \hat{a}_2\hat{c}_1 + \hat{a}_3\hat{d}_1 \qquad (5')$$

Likewise, the "approximate" total effect on property values from *ceteris paribus* changes in school expenditures and municipal expenditures are

$$dV/dE_s = \partial V/\partial E_s + (\partial V/\partial T)(\partial T/\partial E_s) + (\partial V/\partial E_m)(\partial E_m/\partial E_s) \qquad (6)$$

and

$$dV/dE_m = \partial V/\partial E_m + (\partial V/\partial T)(\partial T/\partial E_m) + (\partial V/\partial E_s)(\partial E_s/\partial E_m) \qquad (7)$$

In similar fashion both (6) and (7) may be rewritten in terms of estimated structural coefficients as

$$dV/dE_s = \hat{a}_2 + \hat{a}_1\hat{b}_1 + \hat{a}_3\hat{d}_2, \qquad (6')$$

and

$$dV/dE_m = \hat{a}_3 + \hat{a}_1\hat{b}_2 + \hat{a}_2\hat{c}_2 \qquad (7')$$

Although it is clear that equations (5) -(7) ignore many potential feedback effects, they may nevertheless provide a basis for a judgment regarding the applicability of the Tiebout hypothesis. In short, if the included communities are in Tiebout equilibrium (or nearly so), the estimated effects calculated by equations (5)–(7) should approach zero or at least be considerably smaller in magnitude than the partial effects indicated by observing only the estimates of the structural coefficients (a_1, a_2, and a_3) from equation (1). Results to this effect are presented in the next section.

One final observation regarding equations (5)–(7) may help clarify the preceding discussion. One aspect of Tiebout's original discussion, which has been largely ignored in the subsequent empirical work, is that allocational equilibrium in a given community includes the mix of local services as well as the scale of the local public sector. Accordingly, a comprehensive test for the presence or absence of the Tiebout result would include the effects on property values from marginal adjustments in both the scale and mix of local public services. Equations (5)–(7) provide this twofold test. Equation (5) may be interpreted as the scale adjustment because changes in the tax rate alone imply proportional adjustments in all local public services. Likewise, equations (6) and (7) reflect the effects of mix adjustments because a change in school expenditures or municipal expenditures alone (even with the implied tax change) will result in a marginal change in the mix of local services. It seems clear, therefore, that "full" Tiebout equilibrium requires that the estimates in equations (5)–(7) equal zero.

Empirical Results

The model presented above in equations (1)–(4) was estimated for Oates's sample of northeastern New Jersey suburbs in 1960 and 1970.[6] Each of the four equations is overidentified, and thus estimation of the structural equations can be carried out using Two-Stage Least Squares (2SLS) (Goldberger 1964, pp. 329–36). Table 1 contains these 2SLS estimates for both years. Looking first at equation (1) in both 1960 and 1970 we see that the presence of capitalization of intercommunity differentials in effective property tax rates is indicated by a negative and significant tax coefficient in both years. Likewise, the coefficient estimates of school and municipal expenditures (E_s and E_m, respectively) are positive and significant (except E_s in 1970 which is only significant at the .13 level), also indicating capitalization in both years. By Edel and Sclar's argument, one would conclude from these results that the Tiebout equilibrium apparently has not been achieved in this housing market in either year but that the consistent reduction in the size of all fiscal coefficients from 1960 to 1970 suggests that these communities are closer to the Tiebout result in 1970 than they were in 1960.

On the other hand, when the estimates of the effect on property values from the full budgetary changes derived above in equations (5)–(7) are compared with the single-equation estimates, the results are more indicative of the Tiebout result. Table 2 contains both estimates for 1960 and 1970. In each row of table 2 the single coefficient value is shown first and the full-effect approximations (eqq. [5]–[7]) are shown second. In every comparison shown in table 2 the magnitude of the capitalization effect is smaller when estimated using the full system of equations than

[6] Although Oates's 1960 sample included 53 communities, comparable data for 1970 were available for only 44 municipalities due to the reclassification in 1970 of nine communities designated in 1960 as "urban townships." The estimates displayed in table 1 are for the 44 communities in both years. The estimates for the 53 community sample in 1960 are shown in table A1.

TABLE 1

DETERMINANTS OF RESIDENTIAL PROPERTY VALUES, NORTHEASTERN NEW JERSEY, 1960 AND 1970

INDEPENDENT VARIABLE	1960				1970			
	Eq. (1) [V]	Eq. (2) [T]	Eq. (3) [E_s]	Eq. (4) [E_m]	Eq. (1) [V]	Eq. (2) [T]	Eq. (3) [E_s]	Eq. (4) [E_m]
Constant	−706.98 (−0.243)	−1.705** (−1.964)	252.42** (1.967)	12.64 (0.802)	−1,443.76 (−0.250)	1.415 (1.367)	399.17** (1.953)	−69.24 (−1.589)
T	−3,644.65*** (−4.365)		33.61 (1.056)	21.882*** (4.958)	−1,882.77** (−2.083)		−11.98 (−0.307)	25.538** (2.104)
E_s	26.92*** (3.256)	0.0067*** (3.185)		−0.0777* (−1.755)	10.82 (1.528)	−0.00051 (−0.473)	.	0.0020 (0.042)
E_m	60.12* (1.790)	0.0147*** (2.876)	0.552 (0.666)		40.90** (1.998)	0.0198*** (7.253)	0.701 (0.877)	. .
R	2,419.76*** (3.533)				5,161.23*** (4.130)			.
D	−20.95 (−0.499)				−174.88** (−1.995)			. .
A	−67.623 (−1.244)				−223.41*** (−2.429)			. .
$RVALPR$		−0.00033*** (−4.405)		0.0119*** (4.907)		−0.00019*** (−3.054)		0.0086*** (2.857)
$NRVALPR$		−0.00048*** (−3.615)		0.0099*** (2.972)		−0.00032*** (−6.068)		0.0081** (2.105)
$AIDPP$		−0.0021 (−0.665)	0.0831 (0.208)			0.0075* (1.866)	0.636 (0.817)	. .

PUB		0 0276***	−1.481			0 0114	−2.794	
		(3.228)	(−1 157)			(1.164)	(−1.515)	
RVALPP			0 0027				0 0027**	
			(1.205)				(1.921)	
NRVALPP			0 00091				0 0034	
			(0 395)				(1.060)	
HS			2 624***				7 123***	
			(3 053)				(4.581)	
OWN				−0 477***				−0.982**
				(−3 277)				(−2.378)
DEN				0 00029				−0.0009
				(0 787)				(−0.986)
X				0 334***				0.825
				(2 545)				(1.099)
NWK				−9 571				62.258
				(0 934)				(1.019)
AREA				1 731**				2.803
				(2 243)				(1.445)
SER	1,635 1	0 3408	42 540	7 881	3,032.2	0 4763	81 725	19.565
Corrected R^2	748	494	594	725	746	694	.627	.727
df	37	37	36	34	37	37	36	34

NOTE.—Dependent variables in brackets, *t*-statistics in parentheses below each coefficient, SER = standard error of regression.
* Significant at the 10 level (two-tailed test)
** Significant at the 05 level (two-tailed test)
*** Significant at the 01 level (two-tailed test)

TABLE 2

INTERCOMMUNITY PROPERTY VALUE DIFFERENTIALS ASSOCIATED WITH DIFFERENTIALS IN LOCAL PUBLIC BUDGETS

	1960 ($N = 44$)	1970 ($N = 44$)	1960 ($N = 53$)
$\partial V/\partial T$	−3,644.65	−1,882.77	−4,247.78
dV/dT	−2,329.10	−838.27	1,353.32
$\partial V/\partial E_s$	26.92	0.0	34.97
dV/dE_s	−2.17	0.0	4.27
$\partial V/\partial E_m$	60.12	40.90	91.46
dV/dE_m	6.54	3.62	18.40

NOTE.—In all cases, coefficients not significant at the .10 level or better are assumed to be zero.

when estimated by the single equation. This observed diminution of differential capitalization implies that these communities are closer to the Tiebout result than is otherwise implied. Turning to more specific comparisons, we notice that in both 1960 and 1970 (for the 44-community sample) the negative sign of dV/dT implies that the local public sector is too large. But when all 53 communities are included, dV/dT becomes positive, thus implying that the local public sector is too small.[7] This reversal of signs in response to a change in the number of included communities suggests that caution should be exercised in attempting any explicit conclusions regarding the precise manner in which the scale of the existing budget allocation differs from that which would result in a Tiebout equilibrium. Indeed, the sensitivity of these results to the number of included communities suggests the desirability of additional work in attempting a more precise determination of the geographical extent of single-housing markets of the Tiebout variety.[8]

Concluding Remarks

Given the stature of the Tiebout hypothesis in current discussions of allocative efficiency in the local public sector,[9] Edel and Sclar's insight that intercommunity housing price differentials due to local taxes and expenditure levels do not prove the existence of a Tiebout equilibrium is especially important. It has been the purpose of this paper to take the argument one step further to show that an empirical verification of the

---

[7] The principal cause of this change in sign is that $\hat{c}_1$ ($\partial E_s/\partial T$) is significant in the 53-community case but not in the 44-community sample in 1960.

[8] Straszheim (1975) has made some progress in this direction.

[9] It is important to recognize that the efficiency properties of the Tiebout result is devoid of normative content except when the prevailing distribution of income and wealth is deemed "just." It is not surprising then that critics of our system of local political autonomy would regard favorable arguments concerning the efficiency properties of Tiebout's model with suspicion—thinking that such arguments are really just a facade which protects residents of affluent suburbs from assuming their full share of metropolitan social and fiscal responsibilities.

Tiebout result is more complicated than has heretofore been acknowledged. Finally, it is to be expected that a more definitive answer regarding just how closely communities can be expected to approximate the Tiebout result will be forthcoming when a more elaborate data base containing actual sales data and detailed characteristics of individual housing units and their neighborhoods becomes available.

Appendix

TABLE A1

DETERMINANTS OF RESIDENTIAL PROPERTY VALUES, 53 NORTHEASTERN NEW JERSEY COMMUNITIES, 1960

Independent Variables	Eq (1) [V]	Eq (2) [T]	Eq (3) [E_s]	Eq (4) [E_m]
Constant	−637.68	−0.91	88.75	10.83
	(−0.287)	(−1.401)	(0.895)	(0.601)
T	−4,247.78***		78.83***	31.10***
	(−4.581)		(2.657)	(5.628)
E_s	34.97***	0.0046***		−0.122**
	(4.437)	(3.459)		(−2.216)
E_m	91.46***	0.0172***	−0.299	
	(2.992)	(4.354)	(−0.426)	
R	1,734.36**			
	(2.279)			
D	4.37			
	(0.099)			
A	−78.72			
	(−1.465)			
RVALPR		−0.00022***		0.0141***
		(−4.195)		(4.944)
NRVALPR		−0.00041***		0.0135***
		(−3.691)		(3.612)
AIDPP		0.0015	−0.180	
		(0.553)	(−0.451)	
PUB		0.0178***	−0.049	
		(2.825)	(−0.048)	
RVALPP			0.0059***	
			(3.559)	
NRVALPP			0.0013	
			(0.583)	
HS			1.520**	
			(2.038)	
OWN				−0.655***
				(−3.807)
DEN				−0.0002
				(−0.434)
X				0.264**
				(1.942)
NWK				−0.877
				(−0.078)
AREA				−0.687*
				(−1.754)
SER	1,801.1	0.3054	42.536	10.318
Corrected R^2	.751	.557	.640	.643
df	46	46	45	43

NOTE.—Dependent variables in brackets, t-statistics in parentheses below each coefficient, SER = standard error of regression

* Significant at the .10 level (two-tailed test)

** Significant at the .05 level (two-tailed test)

*** Significant at the .01 level (two-tailed test)

References

Edel, Matthew, and Sclar, Elliott. "Taxes, Spending, and Property Values: Supply Adjustment in a Tiebout-Oates Model." *J.P.E.* 82, no. 5 (September/October 1974): 941–54.

Edelstein, Robert. "The Determinants of Value in the Philadelphia Housing Market: A Case Study of the Main Line 1967–1969." *Rev. Econ. and Statis.* 56 (August 1974). 319–28.

Goldberger, A. S *Econometric Theory*. New York: Wiley, 1964.

Oates, Wallace E. "The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis." *J.P.E.* 77, no. 6 (November/December 1969): 957–71.

Pollakowski, Henry O. "The Effects of Property Taxes and Local Public Spending on Property Values. A Comment and Further Results." *J.P.E.* 81, no. 4 (July/August 1973). 994–1003

Straszheim, Mahlon R. *An Econometric Analysis of the Urban Housing Market*. Urban and Regional Studies, no. 2. New York. Nat. Bur. Econ. Res , 1975.

Tiebout, C. "A Pure Theory of Local Expenditures." *J.P.E.* 64, no. 5 (October 1956): 416–24.

Do Import Levies Matter? The Case of Sweden

Gary Sampson

Monash University

Alexander Yeats

United Nations Conference on Trade and Development

I. Introduction

In recent years considerable attention has been devoted to the evaluation of effective protection afforded by national tariff structures The concept of effective protection is based on the notion that protection of value added, rather than the duty on the product itself, is the primary concern of domestic producers [1] It is argued that, in addition to the tariff on the final product, the influence of duties imposed on raw materials and intermediate goods must also be considered Increasing the duties on material inputs will, *ceteris paribus*, reduce the degree of effective protection enjoyed by a product through the increased cost of producing it.

With few exceptions, previous empirical studies which have attempted to evaluate the magnitude of effective protection have concentrated on tariffs.[2] In part, this may be due to difficulties in deriving empirical measures showing the importance of nontariff barriers (NTBs), or to the assumption that their effect, relative to tariffs, was small. As this paper will demonstrate, neglect of NTBs may have invalidated, or greatly weakened, the conclusions of some previous studies.

This paper concentrates on the incidence, and frequency of use, of one such nontariff restraint. variable levies [3] Levies were studied because they

---

The authors would like to thank Miss Elise Duhy for programming assistance and Prof Odd Gulbrandsen for comments and suggestions The views expressed in this paper do not necessarily reflect those of the authors' institutions.

[1] The discussion in this paper assumes that the reader is familiar with the theory of effective protection For basic references, see Balassa (1965) and Corden (1966).

[2] An important exception is Baldwin (1970) Also, the trade restrictive effects of transport costs and U.S tariffs are compared in Waters (1970)

[3] Essentially, variable levies are fixed charges on imports which are frequently adjusted to maintain equality between domestic and world prices This method of internal price support has been applied since the end of the Second World War by countries such as

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved.

are used by most European countries yet have received little formal attention. An additional factor is that their application is particularly heavy on products of importance to developing countries; however, they also cover items which constitute a high percentage of developed-country exports.[4]

Since there are no previous empirical analyses of the effective protection afforded by levies, it is felt that such a study would be of special interest in the current round of trade negotiations being conducted under the auspices of the General Agreement on Tariffs and Trade (GATT). This analysis will show that the trade-distorting effect of variable levies should be a major consideration for all participants Sweden has been chosen to demonstrate this point since it is frequently cited (erroneously) as having a low protectionist profile [5] However, it will be shown that these conclusions also apply to the European Economic Community

II. The Frequency and Incidence of Swedish Levies

To evaluate the frequency and distribution of variable levies in Sweden, special GATT-UNCTAD (U N Conference on Trade and Development) trade tapes were employed Since these computerized files contain codes for those tariff line items (i e , the narrow product groups used for assessing import duties) which are covered by levies, it was possible to tabulate the number of products and the value of imports covered by these charges Table I presents this information for one- and two-digit Standard International Trade Classification (SITC) groups For comparative purposes, the average tariff rate and average ad valorem equivalent of the variable levies applied to each product group are also given [6]

As indicated in table 1, variable levies are most frequently applied to imports falling in SITC 4 (animal and vegetable oils) and SITC 0 (food and live animals) Some indication as to the importance of these special charges for the LDCs can be derived from the fact that over 40 percent of their total nonpetroleum exports fell in the latter group and close to

---

Austria, Denmark, Finland, Sweden, Switzerland, and the United Kingdom In particular, variable levies form the cornerstone of the Common Agricultural Policy (CAP) of the European Economic Community For a discussion of how this system functions, see UNCTAD (1973).

[4] Out of the 127 products covered in the UNCTAD inventory of nontariff barriers, 41 are exported by, or are of potential export interest to, LDCs. A survey we conducted revealed that over 50 percent of these items are subject to levies in one or more European nations

[5] See, e.g , Balassa (1968) and Healey (1972).

[6] The ad valorem equivalents of the variable levies are for 1969 and were taken from GATT (1971) In some cases, the ad valorem equivalents also include the influence of the Swedish "taxes compensatoir " While the variable levies may be changed frequently, additional data on Swedish intervention prices published in OECD (1974) indicate that the 1969 ad valorem rates are representative of their longer-term levels.

TABLE 1

Comparison of Coverage and Rates of Protection from Swedish Tariffs and Variable Levies

Description and SITC	1970 Value of Imports ($000)	% Coverage by Levies		1970 Estimated Average Nominal Rate (%)	
		Value of Imports	Tariff Line Items	Tariffs	Levies
Food and live animals (0)	627,275	29.5	40.6	3.6	39.8
Meat and preparations (01)	54,970	94.9	79.4	0.0	62.0
Dairy products and eggs (02)	15,471	100.0	100.0	0.0	77.3
Cereals and preparations (04)	33,385	41.2	48.0	6.3	65.9
Sugar, honey and preparations (06)	21,215	62.1	59.3	8.7	105.8
Miscellaneous food preparations (09)	18,590	34.6	35.8	4.1	20.2
Beverages, tobacco, and fuels (1+3)	826,185	0.0	0.0	12.2*	0.0
Crude materials, inedible (2)	614,485	1.8	2.8	0.3	77.3
Animal and vegetable oils (4)	39,855	63.2	50.0	0.2	78.7
Animal oils and fats (41)	8,619	5.6	44.5	0.0	116.2
Fixed vegetable oils (42)	22,885	90.0	77.3	0.1	77.7
Chemicals and plastics (5)	570,154	1.0	2.5	5.2	30.9
Machinery and manufactures (6+7+8)	3,314,553	0.0	0.0	8.9	0.0

Source.—Tariff and trade data from UNCTAD computerized records. Information on variable levies from GATT (1971)

* Excludes fuels

50 percent fell in both groups combined.[7] Variable levies are applied with high frequency in the important meat products group, where 95 percent of imports, measured by value, and approximately 80 percent of the tariff line items are covered. However, import coverage is highest in SITC 02 (dairy products and eggs), where all of the tariff line items must enter under levies.

Perhaps the most striking feature to emerge from table 1 concerns the relative height of the average nominal tariff and average ad valorem equivalent of the variable levies. In SITC 0, for example, the average nominal tariff of 3.6 percent is coupled with a levy of more than *10 times* this figure.[8] In SITC 4, tariffs emerge as being almost negligible, but levies average close to a rate of 80 percent. In summary, the protectionist picture for these products is completely different from that which would be derived from an analysis of nominal tariffs alone.[9]

III. Effective Protection for Selected Industries

While the previous analysis provides an overview of the coverage and importance of variable levies in the structure of Swedish protection, it tells little about the way in which these special charges affect the position of individual industries. Specifically, what is needed is information on how the levies influence the effective rate of protection for different products.[10] To investigate this question, we compiled a list of products for further study. For inclusion, a product had to be either an important export item for a number of LDCs or judged to have the potential to achieve such a position of importance if significant trade barriers were relaxed.[11]

---

[7] These items are also important for developed countries, as approximately 20 percent of OECD-country exports fall into these SITC groups.

[8] Gulbrandsen and Lindbeck (1973) estimate that in 1967 Swedish protection raised agricultural prices 80 percent above world levels. This figure is somewhat higher than we can account for by tariffs and levies (a combined rate of just over 40 percent in SITC 0) and indicates there are other important NTBs restricting trade in these items.

[9] Aside from the high ad valorem incidence on imports, an UNCTAD (1973) study shows that the levy system may restrict trade in other ways. For example, exporters of products subject to variable levies might be adversely affected by the general uncertainty which the system creates. Frequent changes in the levies may cause difficulties for exporters in calculating market prices, information about changes in levy rates may not be available promptly, and, as in the case of the EEC, different interpretations of the existing regulations in member countries as to the raw material content of certain goods might create unexpected problems.

[10] The formula for calculating the effective rate of protection of industry j from both tariffs and levies ($E_{j,TL}$) is

$$E_{j,TL} = \frac{(T_j + L_j) - \Sigma_{i=1}\, a_{ij}(T_i + L_i)}{1 - \Sigma_{i=1}\, a_{ij}},$$

where T_j and L_j show the nominal rate of tariffs and levies on the end product, T_i and L_i show similar information for the inputs, and a_{ij} approximates the free-trade share of input i in the production of j.

[11] Products in this latter category were identified through use of a multiple-regression model which associated skill and other production requirements with availability of these

Once a tentative list was completed, data were gathered on input coefficients for each production process. Here, primary reliance was placed on information contained in the 1963 United States input-output table and similar data drawn from Benelux sources.[12] In addition, extensive use was also made of production coefficients supplied by FAO experts and information published by UNIDO (1971) To complete the data base, tariff information was taken from computerized files available at UNCTAD and GATT The ad valorem equivalents of the levies were gathered from GATT documents While some deletions were necessary where data could not be secured, these sources did provide sufficient information to estimate effective protection rates for 35 key products or product groups

Table 2 summarizes the estimates of effective protection for these products and shows the weighted nominal rate (tariffs and levies) applied to inputs and the final product. To assist in evaluating these results, effective protection estimates derived separately from tariffs and levies are given, as well as those from their combined effects Finally, the table shows the average nominal and effective rates for these products

The primary point that emerges from these results is that estimation of effective protection using tariff data alone would give an erroneous picture of the *level* of Swedish protection for these items This point seems to have been missed in several previous analyses of Sweden's effective protection.[13] For example, meat and dairy products both have a slightly negative effective tariff rate, but the effective protection from tariffs and levies ranges upward to several hundred percent for these items The most extreme cases occur for vegetable oil products, where low value added in processing coupled with high ad valorem equivalents (35 to over 100 percent) for levies sends effective protection to astronomical levels

While the previous analysis demonstrates that examination of nominal or effective tariffs misrepresents the *level* of protection for these industries, a question of equal importance frequently concerns the *distribution* of protection across industries The level of protection reveals the degree to

---

factors in developing countries For full details, see UNCTAD (1970) and GATT (1973). Even with the substantial tariff and nontariff barriers, the 35 selected products (table 2) accounted for 10 percent of Swedish nonfuel imports from LDCs

[12] Recent studies have shown that effective rate estimates should be regarded with caution due to problems associated with inaccurate input coefficients For example, Yeats (in press) has demonstrated that effective protection estimates may be considerably in error when production information pertaining to a previous period is used. Also, Sampson (1974) has shown that the error in input coefficients associated with substitution among factors of production may be significant

[13] For example, an analysis of Sweden's post-Kennedy Round tariff rates estimates zero effective protection for meat products, groundnut oil, coconut oil, and palm kernel oil when the actual rates, due to the imposition of levies on these products, were several hundred percent. In other instances (fish, fruits and vegetables, etc.) levies on basic inputs reduced the effective protection below the rate shown (see UNCTAD 1968) A similar ommission is made in Balassa (1968)

TABLE 2

NOMINAL TARIFFS, VARIABLE LEVIES, AND THE EFFECTIVE RATE OF PROTECTION IN SWEDEN, 1970
SELECTED PRODUCTS OF IMPORTANCE TO DEVELOPING COUNTRIES

PRODUCT NAME	NOMINAL RATE ON INPUTS		NOMINAL RATE ON END PRODUCT		EFFECTIVE PROTECTION		
	Tariffs	Levies	Tariffs	Levies	Tariffs	Levies	Total
Meat products	0.3	9.3	0.0	63.7	−1.0	217.8	216.8
Preserved seafoods	0.6	0.8	4.1	0.1	11.6	−2.3	9.3
Preserved fruits and vegetables	2.6	1.4	13.4	0.0	40.1	−5.3	34.8
Grain and grain products:							
Corn milling	0.2	26.2	0.0	84.3	−0.5	165.8	165.3
Flour and cereal preparations	0.3	36.2	2.9	52.9	13.7	88.0	101.7
Bakery products	0.3	10.0	16.5	0.0	36.0	−22.1	13.9
Prepared feeds	0.2	26.3	0.0	16.0	−1.6	−68.5	−70.1
Prepared and processed foods:							
Animal and marine fats and oils	0.2	1.3	0.0	24.8	−0.9	117.5	116.6
Macaroni and spaghetti	0.5	20.4	54.8	0.0	175.2	−65.9	109.3
Pickles and dressings	2.5	0.0	8.9	0.0	38.8	0.0	38.8
Cocoa powder and butter	0.0	0.0	2.0	0.0	16.8	0.0	16.8
Chocolate products	1.5	2.0	5.2	0.0	12.4	−6.8	5.6
Alcoholic beverages	1.1	6.6	8.7	0.0	15.1	−13.2	1.9
Dairy products:							
Butter	0.1	88.1	0.0	204.0	−1.2	1,158.8	1,157.6
Cheese	0.1	15.2	0.0	39.7	−0.2	178.5	178.3
Condensed and evaporated milk	0.4	66.3	0.0	76.8	−2.1	58.4	56.3

Vegetable oils.							
Soybean oil	0.1	1.3	0.0	104.9	−1.7	1,480.0	1,478.3
Groundnut oil, crude	0.1	0.0	0.0	71.7	−0.7	896.5	895.8
Palm kernel oil, crude	0.1	0.0	0.0	50.8	−0.7	635.2	634.5
Rapeseed oil	0.3	0.0	0.0	92.9	−1.9	617.5	615.6
Coconut oil, crude	0.2	0.0	0.0	56.0	−2.2	509.3	507.1
Cottonseed oil	0.2	3.2	0.0	61.8	−2.0	488.0	486.0
Palm kernel oil, refined	0.2	49.7	0.0	35.8	−6.0	−462.7	−468.7
Tobacco products	0.1	0.0	31.5	0.0	59.2	0.0	59.2
Leather and leather products							
Shoes	2.6	0.0	14.0	0.0	24.5	0.0	24.5
Leather	0.4	0.0	7.3	0.0	23.1	0.0	23.1
Leather manufactures	1.8	0.0	7.0	0.0	12.5	0.0	12.5
Rubber products	0.9	0.0	8.0	0.0	19.7	0.0	19.7
Soaps and detergents	1.1	2.1	8.0	0.0	29.7	−9.1	20.6
Yarns, threads, and fabrics							
Wool fabrics	2.8	0.0	15.3	0.0	39.1	0.0	39.1
Cotton fabrics	1.7	0.0	13.0	0.0	37.7	0.0	37.7
Cotton clothing	6.4	0.0	17.4	0.0	29.7	0.0	29.7
Jute fabrics	0.1	0.0	9.6	0.0	26.4	0.0	26.4
Wool clothing	7.7	0.0	15.0	0.0	19.7	0.0	19.7
Wool yarn	0.2	0.0	5.5	0.0	19.0	0.0	19.0
Average	1.1	10.5	7.7	29.6	19.4	170.2	189.5

the EEC levies was only about 10 percent lower than that for Sweden.[15] These findings suggest that, given their frequency and incidence, considerable attention should be directed toward levies in the current GATT trade negotiations[16]

Finally, these results have implications for the direction of future research aimed at estimation of restrictions to international trade. Since tariffs may now constitute a relatively minor barrier in some countries (such as Sweden), it seems crucial that future research be expanded to account for the influence on nontariff measures. As this analysis has demonstrated, exclusion of the latter may misrepresent a country's true protectionist profile.

References

Balassa, Bela. "Tariff Protection in Industrial Countries: An Evaluation " *J.P.E* 73, no. 6 (December 1965) 573–94.

——— "Tariff Protection in Industrial Nations and Its Effects on the Exports of Processed Goods from Developing Countries." *Canadian J. Econ.* 1 (August 1968). 583–93.

Baldwin, Robert *Nontariff Distortions of International Trade.* Washington · Brookings Inst , 1970

Corden, W. M "The Structure of a Tariff System and the Effective Protection Rate." *J P E.* 74, no. 3 (June 1966) 221–37.

General Agreement on Tariffs and Trade. *Import Measures · Variable Levies and Other Special Charges.* Geneva GATT, April 28, 1971.

——— *Information regarding Import Restrictions on Products of Export Interest to Developing Nations.* Geneva · GATT, 1973.

Gulbrandsen, Odd, and Lindbeck, Assar *The Economics of the Agricultural Sector.* Stockholm. Almquist & Wiksell, 1973.

Healey, D. T. "Development Policy New Thinking about an Interpretation." *J. Econ. Literature* 10 (September 1972): 757–97.

Organization for Economic Co-operation and Development. *Agricultural Policy in Sweden* Paris: OECD, 1974.

Sampson, Gary. "On Factor Substitution and Effective Tariff Rates." *Rev. Econ. Studies* 41 (April 1974) 293–96.

United Nations Conference on Trade and Development. *The Kennedy Round Estimated Effects on Tariff Barriers.* New York · United Nations, 1968.

———. *Trade in Manufactures of Developing Countries: 1969 Review.* New York: United Nations, 1970.

---

[15] This comparison is for 1969 and is based on data presented in GATT (1971). However, it appears that the EEC tariffs on these items are somewhat higher and therefore the combined effect of the two charges is about the same Examination of the UNCTAD inventory of nontariff barriers suggests, however, that the EEC makes somewhat more liberal use of other NTBs on these products than does Sweden.

[16] While it has been alleged that the primary impact of the variable levies is to insulate domestic producers from foreign competition, it should be noted that this system of support can have disruptive effects on world trade flows. For those items in surplus, the CAP provides for disposal of the excess, primarily through export subsidies. Such disposal can displace exports from LDCs in other parts of the world.

———. *Liberalization of Nontariff Barriers.* Geneva: United Nations, June 20, 1973.

United Nations Industrial Development Organization. *Profiles of Manufacturing Establishments.* New York: United Nations, 1971.

Waters, W. G. "Transport Costs, Tariffs, and the Pattern of Industrial Protection." *A.E.R.* 60 (December 1970): 1013-20.

Yeats, Alexander. "An Analysis of the Effect of Production Process Changes on Effective Protection Estimates." *Rev. Econ. and Statis.* (in press).

Money, Interest, and Prices in Market Disequilibrium: A Comment

George H. Blackford

State University of New York College at Buffalo

In a recent article in this journal, Herschel Grossman (1971) has extended Robert Clower's (1965) dual decision hypothesis to the case of multimarket disequilibrium in an attempt to provide a choice-theoretic analysis of the disequilibrium behavior of prices and interest. Two comments are in order regarding this article

First, Grossman (1971, pp 943, 960) claims too much when he argues that his extension of Clower's dual decision hypothesis provides a choice-theoretic foundation for Don Patinkin's (1952) concept of spillover effects. In spite of the assertions "this maximization calculus yields . . ." (Grossman, 1971, p 949) and "he will maximize his objective function . . as follows . . ," (p 950) made in specifying equations (5) and (6) of his model, these two equations are not behavioral relationships implied by a theory of choice They simply *define* the spillover coefficients α_i and β_i in terms of effective and notional demands of his analysis.

Second, Grossman's (p 951) discussion of the effective excess demand for money in his model is inconsistent with the fundamental assumption upon which his model is based Grossman is most explicit on this assumption "The model . . . assumes that the individual always demands . . the quantity which, given his perceived constraints in other markets, he calculates to be optimal, although this optimal quantity may exceed his perceived constraints in its own market" (1971, p 951)

In the paragraph immediately following this statement, however, this statement is ignored: "Finally, in this type of model, as Leijonhufvud (1968, p. 88) has pointed out, the specification of the effective excess demand for money is ambiguous. The effective demands for bonds and commodities, by the budget constraint, imply an effective demand for money equal to $-(b_i' + y_i')$ This implied effective demand can be equal to $-(b_i' + y_i)$ if $y_i' = y_i$, or to $-(b_i + y_i')$ if $b_i' = b_i$, or to neither of these. However, in any case, this implied effective demand in general will differ from both $\pi_i'(\bar{y}_i, b_i')$, given by equation (6.1), and $\pi_i'(y_i', \bar{b}_i)$, given by equation (6.2), and the latter two values will also, in general,

[*Journal of Political Economy*, 1976, vol. 84, no. 4, pt. 1]
© 1976 by The University of Chicago. All rights reserved.

be different."[1] (Grossman 1971, p. 951). *The alleged ambiguity of the effective excess demand for money and the confusion displayed in this paragraph are purely a function of Grossman's inconsistency.* If Grossman's fundamental assumption is applied in specifying the effective excess demand for money in his model, the individual's effective excess demand for money is given (in Grossman's notation) by n_i when he perceives no bond or commodity market constraints on his behavior, by $n_i'(\bar{y}_i, b'_i)$ when he perceives a constraint in the commodity market but not the bond market, by $n_i'(y_i', \bar{b}_i)$ when he perceives a constraint in the bond market but not the commodity market, and by $-(\bar{y}_i + \bar{b}_i)$ when he perceives a constraint in both the bond and commodity markets.[2] Each of these effective excess demands is *the* effective excess demand for money in its respective equilibrium or disequilibrium situation, because *each is implied by the same set of assumptions that imply the other effective excess demands in Grossman's model*

References

Blackford, G "Money and Walras' Law in the General Theory of Market Disequilibrium " *Eastern Econ J.* 2 (January 1975): 1–9

Clower, R. "The Keynesian Counter-Revolution A Theoretical Appraisal." In *The Theory of Interest Rates*, edited by F H Han and F. P. R Brechling. London. Macmillan, 1965 Reprinted in *Readings: Monetary Theory*, edited by R. Clower. Baltimore· Penguin, 1969.

Grossman, H. I "Money, Interest and Prices in Market Disequilibrium " *J.P E* 79, no 5 (September/October 1971) 943–61.

Leijonhufvud, A *On Keynesian Economics and the Economics of Keynes*. New York. Oxford Univ Press, 1968

Patinkin, D. "The Limitations of Samuelson's 'Correspondence Principle'." *Metroeconometrica* 4 (August 1952) 37–43

[1] Grossman's reference to Leijonhufvud in this paragraph is puzzling The distinction made by Leijonhufvud (*1968, pp 87–89*) *is one of semantics*—"a matter of terminology"—and not of substance Leijonhufvud's discussion does not justify the assertion that the effective excess demand for money is ambiguous in disequilibrium models (cf Clower 1969, p 295, n)

[2] In Grossman's three-good model the individual has zero degrees of freedom in "choosing" his effective excess demand for money when he perceives constraints in both the bond and commodity markets As a result, his effective excess demand for money is given by substituting his perceived market constraints into the budget constraint For a more complete discussion of the effective demand for money within the context of general disequilibrium analysis, see Blackford (1975)

Book Reviews

Free Trade and Frustration. Anglo-Austrian Negotiations, 1860–1870 By KARL F. HELLEINER.

Toronto and Buffalo. University of Toronto Press, 1973 Pp. vii + 152.

Economists are infuriated or intrigued, according to their temperaments, by the fact that governments entrusted with tariff policy persist in viewing their own tariff cuts as a loss and those of their opponents (i e., partners) as a gain. This view of trade policy makes tariff negotiations akin to disarmament talks, in which the protagonists confront one another, each demanding that his opponent relinquish what he is most impressed by in the latter's armory, each anxious to preserve what he most prizes in his own arsenal Negotiations can only be successful if opinions differ on both sides as to what constitutes each negotiator's principal source of strength. With such a coincidence of mutually compatible delusions, each party can convince itself that it has given up less than it has obtained in the course of the confrontation. On the other hand, if each side clear-sightedly wants the other to relinquish what he most desires to retain, the prospects for disarmament, tariff or otherwise, are dim Since in trade matters governments tend to protect precisely those industries which shout the loudest that they are about to succumb to import competition and wield the heaviest political clout because of their size, and since these industries' markets represent the largest and most desirable negotiating objectives of foreign governments in the event of a trade negotiation, trade talks tend to avoid the areas where real gains from specialization and trade could be made and to concentrate on "negotiable" issues, namely, those about which neither side feels very strongly This is what I should like to call the most-trivial-issue principle, upon which much commercial diplomacy is based.

Professor Karl F. Helleiner has written with loving care of the commercial negotiations between Britain and Austria which began in 1860 and, after 10 years of labor, gave birth to a mouse of an agreement

The book starts with the British overtures in 1860 aimed at reducing an exorbitant Austrian export duty on rags and ends with the modest reduction in 1869 of the Austrian import duty on cottons and woolens—a classical example of the most-trivial-issue principle outlined above. In between these dates Britain made a dogged effort to negotiate a commercial treaty with Austria on the free-trade, most-favored-nation model which had just been successfully concluded with Napoleon III. These efforts were temporarily rewarded in 1865 with an agreement which reduced Austrian tariffs across the board in exchange for access to the British market on the same terms as those accorded to France, with an added sweetener in the form of lighter fiscal duties on bottled wine Scarcely was the ink dry on this agreement, however, than the Austro-Prussian War of 1866 broke out. In the ensuing renegotiation of Austria's relationship with Prussia and the Zollverein, the Austrian tariff reductions specified in the Austro-British commercial treaty of 1865 were postponed on the grounds that Austria could not negotiate with the Zollverein (which she had wanted to join on her own protectionist terms for years) shorn of all her negotiable tariffs.

Negotiations on the application of the 1865 treaty resumed in 1867 and dragged on until the end of 1869, by which time the British negotiators gave up all hope of getting anything more out of the Austrians than the latter had already given to the French and to the Germans, for, as Helleiner points out, the British had, in the meantime, given away most of their bargaining counters by moving to free trade unilaterally. Thus did Austria successfully thwart Britain's efforts to drag it into the free-trade camp. It is just one of the ironies of history that barely a century later the free-trade shoe was on the other foot, and Austria, in company with the other neutral countries of Europe, succeeded in dragging a reluctant Britain into the European Free Trade Association.

The study reminds us that the British government of the period took an economist's view of trade policy and preached "the glad tidings of Free Trade to the children of Protectionist darkness," as Helleiner puts it so nicely. They were confident that, once the Continental powers had understood the free-trade argument, they would embrace it unilaterally In this they were doomed to disappointment, for the very same political forces which in Britain had led to the country's conversion to free trade, on the Continent reinforced the traditional forces of protection. There were good political reasons for Britain's missionary zeal once the landed gentry and aristocracy had been compensated for their loss of agricultural protection due to the repeal of the Corn Laws in 1846 with the challenging job opportunities offered by an expanding empire, there were virtually no forces of protection to contend with in Britain British industry feared no inroads by foreigners in its own home market and was in fact eager to meet such competition as there was head on in foreign markets—provided the government could negotiate suitable tariff terms of entry It thus made good political sense for the British government to attempt to spread the free-trade gospel, just as it made good sense for the United States to promote freer trade after World War II, and it is only logical that Japan should today be playing godparent to the current round of trade negotiations in GATT (What will happen when there is no one left to pick up the torch but Hong Kong is only too predictable.)

What was far harder to accomplish, and ultimately proved impossible, was to convince the children of protectionist darkness that what was good for Britain in 1860 (or the United States in 1945, or Japan in 1975) was good for them, too. Even if they could be intellectually convinced of the argument that they would be better off if they got rid of their more inefficient industries, they could not overcome the political opposition of the nascent manufacturing class, fearful of British competition, convinced of its merit, and increasingly powerful in political terms.

Thus, the rise of the bourgeoisie in nineteenth-century Europe gave strength to free-trade forces in Britain and nurtured protectionist forces elsewhere, and Helleiner notes sadly that "it was experiments with parliamentary institutions that created the most serious obstacles for a policy of economic liberalism in Austria." This is indeed a damning and thought-provoking comment. Are parliamentary democracy and economic liberalism really so incompatible? Is this the reason why parliamentary democracies to this day take a mercantilistic view of trade negotiations? What constitutional changes could be advocated which would limit the power of vested industrial interests and enhance those of the consuming and voting population?

It is a sobering thought that, although the subject matter of the book is over 100 years old, it abounds in revealing parallels with contemporary trade policy issues. Helleiner's scholarly investigation is full of new material based on hitherto unpublished British and Austrian sources and is organized into seven brief and readable chapters. Students of trade policy and economic history will remain in Helleiner's debt for the light he throws on this little-known but instructive episode in nineteenth-century commercial diplomacy.

VICTORIA CURZON

Institut Universitaire d'Etudes Européennes,
Center for Education in International Management,
Geneva University

Trade Instability and Economic Development An Empirical Study. By ODIN KNUDSEN and ANDREW PARNES
Lexington, Mass. · D. C Heath & Co., 1975. Pp. x + 143.

Ever since Coppock's major statistical attack on international economic instability, there has been a growing literature on the measurement, causes, and consequences of export instability, with special reference to the less-developed countries. The book under review is the latest addition to the literature. The issues considered in the literature are whether there has been a higher degree of export instability for less-developed countries than for developed countries; whether the trade in primary commodities has been more unstable than that in manufactures; the relation between instability in export earnings on the one hand and price and quantity fluctuations on the other, the role of disturbances on the demand side and on the supply side, how far export instability has been due to such factors as the proportion of primary products in a country's exports and the commodity composition and geographic distribution of its exports; and whether a high degree of export instability has had adverse effects on a country's investment, rate of growth, and overall stability. This literature is briefly reviewed in the first quarter of the book. So far, the conclusions have been confusing and contradictory, and there has been a need for some time for a definitive treatment of the data and the analysis, but it is not to be found in the present book.

The way the game is played is for each writer to invent a new measure of export instability, apply it to a group of countries for which data are conveniently available to him, fit a series of regression equations, and announce his results. The present book is no exception, in fact, it is based on two new approaches to the measurement of export instability The first measure is presented in chapters 4–6, the novelty being that an index is calculated for the trade of each country with each of its trading partners for the period 1954–67. Analysis-of-covariance methods are then used to test whether the ratio of corrected mean squares of these indices between countries to that within countries is statistically significant. It is then found that "the choice of trading partners has an important influence on the level of instability" and that "by not accounting for these country-specific effects, previous studies have overlooked an important factor in explaining variations in export instability among nations" (p 129) However, it must be said that this approach and that of earlier writers address themselves to different questions. By starting from the trade with individual partners, this approach leads to a different concept of instability, one whose link with total export instability considered by earlier writers is not made clear. In some applications, the authors have gone some way toward this link by classifying their indices by countries of origin or of destination and regressing on a number of variables such as commodity concentration, percentage of food exports, percentage of raw materials exports, volume of exports, and level of development of trading partner. In this way, the authors find that export instability is significantly increased by greater commodity concentration, by smaller volume of exports, and by trade with less-developed countries. Insofar as total export instability is concerned, the gain is not much. As far as their measure of country-specific instability is concerned, it might have been more useful if the authors had presented their results of the effect of individual countries on the export instability of the countries from which they import.

Their second measure of export instability, presented in chapters 7–9, is based on the permanent-income hypothesis of consumption. An adaptive-expectations model is used to separate permanent from transitory components of both domestic and export incomes, and the variances of the transitory components are then used as measures of instability. Estimates are derived for 28 less-developed countries, of which 16 are from Latin America, surely a rather biased sample. Like other measures of instability, the new measure is based on divergences of actual export earnings from projected values, in estimations of which the trend of past values enters quite significantly; it is therefore not surprising that the new measure is quite highly correlated with other measures. However, the authors say that "the contribution of this approach does not depend on the definition of yet another new, more refined index of instability" (p. 131). Instead, their main aim is to estimate the effect of the uncertainty caused by income instability on behavior affecting economic growth. To this end, they use statistical estimates of the marginal propensity to consume out of the permanent components of income, derived in the process of separating these components from the transitory components. Their most interesting result comes from their regression of the marginal propensities to consume on their indices of instability. They find that, while the regression of the marginal propensity to consume out of permanent domestic income on instability of domestic income is statistically significant at a 5 percent level, the regression of the marginal propensity to consume out of permanent export income on instability of export income is only significant at the 10 percent level. In addition, they regress investment and the growth rates of total and per capita GNP on an average of their indices of instability of domestic and export incomes and find positive regression coefficients. As one of the issues which started the whole discussion was whether investment was related to export instability, the authors might have presented their results for export instability separately. In the one case where they do (p. 122 n.) the significance level declines to 10 percent in a one-sided test, but on the whole the authors stay away from the subject with the argument that "aggregate investment cannot be explained by just export instability or just domestic instability" (p. 122).

On the strength of these results, they come to the conclusion that "instability does indeed stimulate certain economic parameters in a direction predicted by the permanent income theory" (p. 131). Their argument is that "instability reduces the propensity to consume by requiring larger reserves for unexpected or temporary declines in income. The higher rate of savings, which results from the lower propensity to consume, makes possible higher investment expenditures. Investment, in turn, leads to higher rates of economic growth" (pp. 131–32). Having lavished so much statistical effort on just one link in this long chain of argument, they might have spared some attention for some of the other links, for example, by examining whether reserves did vary in the manner assumed or whether such changes had the effects postulated. An aspect of the problem neglected by these authors and also others who have reached similar conclusions is the tendency for instability of exports to be positively correlated with rates of growth of exports, a tendency to be found in a number of other fields as well. This raises the question, worth further consideration, of whether the positive relationship found between export instability and investment rates might not really be due to the higher growth rates of exports associated with higher export instability.

The logic of the argument the authors set out to demolish is that, if a country were planning its investment program tightly over the medium term, and if it relied heavily on export earnings to finance the foreign exchange costs of its investment program, then this program would be disrupted by export instability with adverse consequences on its economic growth. This, at least, was the basis of the UK-Swedish proposal to UNCTAD I for Supplementary Financial Measures. If the object of the exercise was to challenge this logic, then perhaps it might have been more useful to study the planning processes in individual countries in more detail than to engage in such contorted statistical exercises on data from a motley sample of countries.

R. M. Sundrum

Australian National University

Classical Economics Reconsidered. By Thomas Sowell
Princeton, N J.: Princeton University Press, 1974. Pp 152.

This book is one of three that have appeared within the last year or so on the same topic The other works are in the nature of a technical monograph (R. V. Eagly, *The Structure of Classical Economic Theory* [1974]) and a general textbook (D. P O'Brien, *The Classical Economists* [1975]) There are no indications in the present work regarding intended readership.

We consider first the scope of the book The Keynesian and Marxian definitions of "classical" economics are rejected as "tendentious," "egocentric," and "idiosyncratic." In their place we read at the outset (pp 4–8) of a tradition established in the *Wealth of Nations* and lasting until the marginalist revolution of the 1870s. The major contributors are Ricardo and the younger Mill as well as Smith and, in the second rank, McCulloch and James Mill There were also those who "contributed key concepts to classical economics without sharing all its methods and conclusions" (such as J B. Say, Malthus, West) and those who were "classical in some respects but not in others" (Torrens, Senior, Marx) It would appear therefore to be the author's view that there existed a "small solid core" sharing a common set of philosophical presuppositions and common methods and conclusions relating to matters of substantive economic analysis The major propositions listed include the labor theory of value, Say's Law, the Malthusian population theory, and the quantity theory of money, while throughout (pp. 33, 74) it is emphasized that the orientation of the classicals in their theoretical and policy writings was toward the issue of economic growth Professor Sowell describes most effectively some of the features common to the group in his discussions of the breakaway by Smith from mercantilist conceptions relating to the nature of wealth, international commercial relations, and the scope of the "nation" (pp. 8ff., 54).

Yet for all that the general impression left by the work as a whole is one of much less coherence than implied above. This is true of many of the basic theoretical positions. To define a clear-cut "classical" view of Say's Law turns out to be all but impossible. Most striking of all is the contrast between Smithian and Ricardian method which emerges in the final chapter on methodology. Smith's approach is represented as "eclectic," entailing an intermingling of "the empirical, the theoretical, the institutional, the philosophical, the static and the dynamic." With Ricardo, we are told, "economics took a major step toward abstract models, rigid and artificial definitions, syllogistic reasoning—and the direct application of the results to policy. The historical, the institutional, and the empirical faded into the background, and explicit social philosophy shrank to a few passing remarks. Comparative statics became the dominant—though usually implicit—approach" (pp. 113–14). Reference is made (p. 121) to classical economics in its "contemporary Ricardian form," and in fact the final chapter is almost entirely devoted to contemporary and subsequent criticisms of Ricardo and "the Ricardians," with very great weight placed on Schumpeter's charge relating to the so-called Ricardian Vice (pp 122, 135–36, 146–47). On the author's own reading of the record, therefore, there were profound modifications introduced into political economy in 1817. We are obliged to raise the question whether any good purpose is served by thinking in terms of a common tradition commencing in 1776 and lasting for approximately 100 years delineated by the term "classical" economics.

We turn now to the substance of the argument. In the first chapter ("Social Philosophy"), a number of interpretations of the "classics" are turned down in rather spirited terms. Attention is paid in particular to the contention that they were apologetic defenders of the status quo, that they believed in a harmony of class interests, and even that they were motivated by personal interest (pp. 8 ff., 17, 26 ff) I am totally in sympathy with Professor Sowell's position as far as concerns Smith, Ricardo, Mill, and others to whose defense he comes. Nonetheless, it seems unnecessary to fight the battle with such energy. One still meets occasionally the distorted picture of a group of hard-hearted grinders of the faces of the poor, but surely the air has been sufficiently cleared—for example, by Lord Robbins's *Theory of Economic Policy*. This impression may reflect wishful thinking, but the author gives not a single reference to the secondary literature to demonstrate the need for a new refutation.

The core of the work as far as concerns analytical matters is contained in two chapters entitled "Macroeconomics" and "Microeconomics," respectively. It is not certain whether much is gained by this particular division of the materials which reflects our modern tastes To some extent there is a net loss in comprehension. It is, for example, difficult to appreciate fully the discussion of limits to growth treated in the first of the two chapters without taking into account the effect of accumulation upon the profit rate in consequence of diminishing agricultural returns, a matter considered only in the second chapter. The discussion of the Malthusian population theory in a chapter on microeconomics also seems poorly placed.

There is much that is valuable in the analytical sections. Their merit lies in an attempt to spell out the ambiguities of various key theoretical conceptions, to fit the argumentation into a suitable frame of reference reflecting contemporary preoccupations, and to demonstrate how confusion was sown by terminological and methodological differences between disputants—and above all by Ricardo's ascriptions to his opponents of his own, largely comparative-statics approach.

Sowell's discussion of Say's Law, which constitutes the essence of the first of the two theoretical chapters, is particularly worthwhile. As in his well-known *Say's Law: An Historical Analysis* (1972), the author takes issue with those who believe that the great controversy turned upon the question of secular stagnation. Both the general-glut and the orthodox economists, he insists, accepted that there were no permanent or secular growth constraints, and the debate turned entirely upon the possibility of short-run excess supply. Thus while Malthus, for example, maintained that increments of saving (investment) might fail to yield an adequate return and would be followed by subsequent disinvestment, the Ricardians insisted that there was no such thing as an equilibrium level of aggregate output (pp. 42 ff.).

Sowell's position is a little difficult to grasp in the light of a summary statement to the effect that "Say's Law answered those who feared that economic growth had reached, or was approaching, some ultimate limit to what the economy could profitably absorb" (p. 72). Since he does not here have in mind the dissenters from Say's Law it would have been most desirable to give the reader some idea of the identity of those contemporaries who did fear secular stagnation. Second, Malthus et al., on Sowell's own showing (pp. 50–51, 72), recognized that further advance beyond the short-run limit to sustainable investment hinged upon technological innovations, but little is said about the prospects for such improvement in their writings. We are told also that J. S. Mill developed a position on oversaving similar to that of Malthus, Lauderdale, and Chalmers (pp 50, 70), but full justice is not done to Mill, since the question of the effects of accumulation on the profit rate by way of the rising cost of wage goods, which played a large part in his work, is not discussed in the present context but, as already noted, in the chapter on microtheory.

A most important feature of the discussion is the emphasis upon Say's conversion in the fifth edition of the *Traité* (1826) regarding the possibility of short-run limits to production and his acceptance of Sismondi's particular version of aggregate output equilibrium. The recantation, Sowell observes, "made no impression on the British classical school" (p 48) and has been neglected by modern commentators.

Sowell insists that none of the major general-glut economists based his position on leakages from the income flow (p 45) But although this is a common view, it probably requires some modification Say himself on at least one occasion recognized deficient money demand as a cause of contemporary depression. (A famous footnote to this effect in Say's *Letters to Mr Malthus* [1821] is mentioned no less than four times in Sowell's book [pp 50, 56, 61, 65]) Now while Ricardo objected that Say "concedes too much," Malthus himself was delighted observing that "he fully concedes all that I contend for" (Sraffa 1951–73, 8 260, 267) There is at least an element in Malthus's position in his debate with Ricardo turning upon the possibility of leakages.

As mentioned earlier, the labor theory is listed among the major propositions of classical economics at the outset of the book But we read, in the chapter on microeconomics, that "the substantive theories of [the classical] system can be presented without the so-called labor theory of value" (p. 75), and in the final chapter that "for all the controversy generated by the 'labor theory of value' it was tangential to classical value theory . . " (p 110) There is no doubt that Ricardo, for one, recognized numerous qualifications to the labor theory of exchange value, but it is also true that he used the theory consistently (indeed more so in the third than the first edition of his *Principles*) in the derivation of his major results. It is possible to restate his conclusions in other terms, but the flavor of the argument is diluted in consequence. Sowell seems here to have neglected one of his own rules of desirable procedure.

The chapter on microeconomics gives what is now the standard view of Ricardian profit theory, namely, that version which runs in terms of the subsistence wage, based upon the Malthusian population principle, combined with the law of diminishing agricultural returns. The notion of profit-rate decline is presented not merely as an analytical but also as an historical proposition (pp. 83–85).

This is a work which can be warmly recommended to the serious student. But it would have been even more useful as a text had the author allowed himself considerably more space to expand his position and relate it to the voluminous secondary literature which now exists.

SAMUEL HOLLANDER

University of Toronto

Reference

Sraffa, P., ed. *Works and Correspondence of David Ricardo*. 11 vols. Cambridge: Cambridge Univ. Press, 1951–73.

The Cobb-Douglas Production Function Once Again: Its History, Its Testing, and Some New Empirical Values

Paul H. Douglas

Washington, D C

Editors' Note—We are pleased to have been offered, and to offer the readers of the *Journal*, this account of the early development of empirical work on the Cobb-Douglas function by its coinventor and pioneer economic investigator, Paul H Douglas

Research into the production function has a long history. Since the first work, in 1928, many studies have tended to support the hypothesis that production processes are well described by a linear homogeneous function with an elasticity of substitution of one between factors New results are presented here, using 7 years of observations on Australian manufacturing industries during the 1950s and 1960s In all seven cases, constant returns to scale are very closely approximated, and the coefficient for labor hovers near 0 6 The appropriate coincidence of the estimated coefficients with the shares received strengthens the competitive theory of distribution

This paper is an effort to continue and extend earlier studies of the production function that were first begun nearly a half century ago. For it was in 1927 that I computed the index numbers of the total number of manual workers (L) employed in American manufacturing by years from 1899 to 1922, did the same for fixed capital (C), expressed these in logarithmic terms on a chart, and then added the index for physical

---

Work on the production function was carried on by a large number of persons who deserve to be credited as coauthors Foremost among these is Grace Gunn who participated, over a period of 40 years, in no less than three separate phases of the work Mary Hook helped in the later stages and Stanley Horowitz and Christopher Jehn in the final preparation of this article I am indebted to all of them and many others for their help Others who helped in the production were Marjorie Handsaker, Martin Bronfenbrenner, and Patricia Daly

[*Journal of Political Economy*, 1976, vol 84, no 5]
© 1976 by The University of Chicago. All rights reserved.

production (P) in manufacturing.[1] I found the curve for product to lie, in general, approximately one-quarter of the distance between the curve for labor, which had increased the least (to 162), and the curve of capital, which had increased the most (to 431). The year 1899 was taken as 100. I was then temporarily lecturing at Amherst College, and consulted with my friend and colleague, Charles W. Cobb, a mathematician. At the latter's suggestion, the formula $P = bL^kC^{1-k}$ was adopted, a form that had also been used by Wicksteed and Wicksell. This, following Euler, was a simple homogeneous function of the first degree. After finding the value of k by the method of least squares to be .75, we found that the estimated values of P closely approximated the actual values for the 23-year period—with such differences as occurred being due primarily to the business cycle (Cobb and Douglas 1928, pp. 139–65). The National Bureau of Economic Research had found that the average share of products going to labor during the decade 1909–18 was 74.1 percent,[2] or an almost precise agreement with the values of the production function. I felt, therefore, that the marginal productivity theory of wages had received a substantial degree of confirmation.

Later studies by Cobb for Massachusetts, 1890–1928 (Cobb 1930), and by Director for New South Wales, 1901–27 (unpublished), gave values which were identical in the case of Massachusetts ($k = .743$) and similar in the case of New South Wales ($k = .65$).

We made one more study of a time series. In 1936, in cooperation with Mrs. Marjorie Handsaker, series of P, L, and C were worked out for the years 1907–29 for the state of Victoria in Australia, and the value of k was found to be .71. Labor's share of the product W/P was computed as .61 for this period (Handsaker and Douglas 1937 and 1938).

Research between 1937 and 1947

With the David Durand article (1937) and the coming of Grace Gunn to the staff in the same year, several changes were made. First, following the suggestion by Durand, the formula $P = bL^kC^{1-k}$ was changed to $P = bL^kC^j$, thus making the exponent of C independently determined instead of treating it as the residual in a homogeneous linear equation. With j independently determined, the production function was no longer constrained to be homogeneous of degree 1, but could instead take such form as the actual figures might dictate. If $k + j = 1.0$, the economic system was subject to constant returns to scale. If $k + j$ was greater than 1.0, then a 1 percent increase in both L and C would be accompanied by an increase of more than 1 percent in product, and the system as a whole would operate under increasing returns. If $k + j$ was less than 1.0, then the system was characterized by diminishing returns. We corrected the time studies according to the Durand formula and found, interestingly enough, that the sum of $k + j$ was very close to the previous assumption of unity. The values of k were reduced to close to two-thirds, that is, .65, and j to approximately one-third At the same time, the National Bureau of Economic Research was revising its previous studies on the distribution of product in manufacturing and had independently arrived at approximately 65 percent as labor's actual share The two independent studies of production and of distribution were, therefore, once again in unison.

---

[1] This index had been developed by E. E. Day.

[2] See Mitchell et al. 1922, p. 98.

The second important change introduced was to substitute cross-section studies of separate industry observations for the previous time series. The use of time-series data in computing the production functions carried with it a host of technical problems, mainly resulting from the necessity of calculating comparable index numbers for capital and for physical product; the necessary adjustments were complicated and time consuming and, thus, severely limited the number of yearly observations available for estimating the function There was also the possibility that the true values of the exponents changed over time

All this was noted by such critics of the production function analysis as Horst Mendershausen and his mentor, Ragnar Frisch They urged that so few observations were involved that any mathematical relationship was purely accidental and not causal They believed sincerely that the analysis should be abandoned and, in the words of Mendershausen, that all past work should be torn up and consigned to the wastepaper basket

This was also the general sentiment among senior American economists, and nowhere was it held more strongly than among my colleagues at the University of Chicago I must admit that I was discouraged by this criticism and thought of giving up the effort, but there was something which told me that I should hold on Miss Gunn and I, therefore, determined instead to broaden the scope of our studies and to take as our observations individual industries within a given economy in a given year (Bronfenbrenner and Douglas 1939, Gunn and Douglas 1941, 1942; Daly, Olson, and Douglas 1943) This methodology eliminated the problem of changes in the production function over time, but it had to assume identical functions across industries

We found a wealth of material The American census had computed data on P, L, and C for 1889, 1899, and every 5 years thereafter Finally in 1931 the series was put to death by an official advisory commission of eminent economists and statisticians Not improvement but decapitation was their motto. We determined, however, to utilize the existing material and worked with several collaborators We also found abundant statistics on P, L, and C in the annual censuses of manufacturing for the various Australian states and for the commonwealth as a whole. These were

TABLE 1

PRODUCTION FUNCTION BASED ON AMERICAN CROSS-SECTION STUDIES, 1904, 1909, 1914, 1919

Year	k	SE of k	j	SE of j	$k + j$
1904	.65	.02	.31	.02	.96
1909	63	02	.34	.02	.97
1914	61	.03	.37	.02	.98
1919	76	.02	25	02	1 01
Average	66	.02	.32	.02	.98

started by the great Australian statistician G H Knibbs, who deserves to rank with Carroll D Wright and R H. Coats in the front rank of the applied statisticians of the century

We had time to finish only four cross-section American studies (1904, 1909, 1914, and 1919) involving 1,490 observations when the war intervened Gunn then entered the governmental statistical service, I enlisted in the marines Articles were published giving the results, but in the excitement caused by the war they did not attract much attention The results, however, were notable (table 1)

In the first place it should be noted that the sum of the independently computed exponents was very close to unity, averaging 98 or only 2 percent less than true constant returns The appropriateness of the previous k and $1 - k$ function was, therefore, reinforced as a first approximation

The relative smallness of the standard errors of estimate gave a further corroboration to the equation. The standard errors of both k and j were found to be only 02 This was only one thirty-third the exponent of L and one-sixteenth the exponent of C Another of the criticisms of Frisch and Mendershausen, therefore, was still further weakened

Work was stopped for some years because of the war and its aftermath. In 1947, however, Gunn and I joined forces again. I had been elected president of the American Economic Association and determined to write my address on the production function covering both the theory and an empirical analysis of the problem. Gunn obtained a leave from the government And so, 20 years after the initial Cobb-Douglas effort, we continued on our course. More cross-section studies, both for the United States and especially for Australia, were launched The years 1889 and 1899 with 695 observations in all were added in the American study, making with the pre-1942 analysis a total based on 2,185 industry observations The values of k for 1889 and 1899 were .51 and .62 and for j .43 and .33. The sum of the exponents was, therefore, .94 and .95. Since the probable errors for k were .03 and .02, and for j were the same, the results strongly indicated at least an approach to true constant returns

TABLE 2

PRODUCTION FUNCTION FOR AUSTRALIA,
SELECTED FISCAL YEARS

Cross-Section Studies and Fiscal Year	Observations (N)	Values of k	SE of k	Values of j	SE of j	$k + j$
Australia:						
1913	85	.52	.05	47	.05	.99
1923	87	53	.05	49	05	1.02
1927	85	59	05	34	.04	.93
1935	138	64	04	36	.04	1.00
1937	87	49	04	49	.04	.98
Victoria:						
1911	34	74	08	25	.11	.99
1924	38	62	08	31	10	.93
1928	35	59	07	27	09	.86
New South Wales						
1934	125	64	04	.34	03	.99
Average of all commonwealth and state studies	714	60	06	37	.06	.97
Average of commonwealth studies only	482	55	04	43	04	98
Average of state studies only	232	65	.07	29	08	.94

but with a continuing tendency during the third of a century for the sum of $k + j$ to be slightly less than unity

Australian Commonwealth cross-section studies were completed for fiscal years 1913, 1923, 1927, 1935, and 1937, while we also covered the state of Victoria for fiscal years 1911, 1924, and 1928, and New South Wales for fiscal 1934 (table 2)

An analysis of these figures indicates the following

1 The sum of the exponents for labor and capital closely approached unity, or true constant returns, but were nearly always slightly below unity

2 The standard errors of k were comparatively slight, ranging from one-tenth to one-fifteenth the value of k The standard errors of j for the commonwealth and New South Wales were also relatively small, while those for Victoria for both k and j were much larger

3. The average exponent of labor, or k, for all nine Australian studies was .60, and the exponent for capital, or j, was .37. The k's for the commonwealth studies only were somewhat lower (average of 55) and the j's somewhat higher (average of 43) The averages in the four state studies were in reverse. That of the k's was higher (average of 65) and of the j's lower (average of 29)

A supplementary study by George Brinigar and Keith Campbell for Queensland in 1937–38 found k to have a value of .58 and j one of .45.

If we take .60 as the most probably "normal" value of k in Australia, and .37 for j, the elasticity of the marginal productivity curve for labor was somewhere around 2.7 and for capital about 1.7.

But we did not stop with Australia. We pushed on to include Canada, New Zealand, and South Africa. We carried out four cross-section studies for Canada covering 1923, 1927, 1935, and 1937 which included a total of 659 industry observations (Daly and Douglas 1943). In addition, G. W. G. Browne made two cross-section studies for South Africa in 1937–38, including in all 102 industry observations (Browne 1943), J. W. Williams covered 61 industries in his cross-section study of New Zealand manufacturing in 1938–39 (Williams and Douglas 1945); Max Brown, in his unpublished doctoral dissertation at Cambridge University, covered New Zealand manufacturing during the 18 years from 1915–16 to 1935–36 (the war year 1917–18 was omitted); J. W. Williams also covered 18 years from 1923 to 1940; K. S. Lomax studied the British economy in 1924 and 1930 (Lomax 1950); and C. E V. Leser studied British coal mining between 1943 and 1953 (Leser 1955). The results for the South African, Canadian, New Zealand, and U K cross-section studies are shown in table 3.

In South Africa the value of k when blacks and whites were combined in 17 industries was 66, when separated the value of k was .65. The j's were, respectively, 32 and 37, or almost identical values They were, moreover, very similar to the k's and j's which had been discovered for American manufacturing and were not greatly different from the averages for all the Australian studies The values of k were, however, consistently lower and the j's higher in Canada than in the United States and Australia The k's ranged from 43 to 50 averaging 47, while the j's ran from 48 to 58 averaging 52

The New Zealand cross section of industry observations also indicated a j which was higher than its k (i e., $k = .46, j = 51$) This was also true of the $P = bL^kC^j$ formula when used for the time series When the $P = bL^kC^{1-k}$ formula was used, however, the k was higher, being 51 for 1915–35 and .54 for 1923–40

Surveying this mixed group as a whole, we were reassured to find that the sum of the exponents of $k + j$ approximated unity, but was very slightly under 1 0. The studies by Lomax were least consistent with a finding of constant returns to scale, but they are difficult to evaluate since no standard errors were reported

The standard errors of k in the four cross-section Canadian studies were all 04, or from one-eleventh to one-twelfth the values of k. The standard errors of j were also uniformly .04, or from one-twelfth to one-thirteenth the values of j. The standard errors in the case of the two South African studies were somewhat higher, amounting in all three cases to .08. This was one-eighth the value of k and one-fourth to one-fifth the value of j.

TABLE 3

PRODUCTION FUNCTION FOR OTHER BRITISH COMMONWEALTH COUNTRIES

Country and Year of Interindustry or Cross-Section Studies	Industry Observations (*N*)	Value of *k*	SE of *k*	Value of *j*	SE of *j*	Value of *k* + *j*
South Africa:						
1937–38*	17	.66	.08	.32	.08	.98
1937–38†	85	.65		37	.08	1.02
Canada:						
1923	167	48	04	48	.04	.96
1927	163	46	04	52	.04	.98
1935	165	50	.04	.52	.04	1.02
1937	164	43	04	58	.04	1 01
New Zealand:						
1938–39‡	61	46		51		.97
United Kingdom						
1924§		.72		18		.90
1930§		73		13		.86
1943–53‖	88#	42	13	60	08	1.02
1943–53‖	99#	51	12	49	08	1 00

* Study by Browne
† Study by Browne with whites and blacks separated
‡ Study by Williams
§ Study by Lomax
‖ Study by Leser
# This study combined time series and regional cross-section observations from British coal mining

Another test of the consistency of the hypothesized relationship calls for an examination of the residuals from the estimated production function.[3] We would expect positive residuals in industries characterized by (*a*) monopoly or highly imperfect competition or (*b*) expanding demand where the demand curve as a whole was shifting to the right. Such industries would have higher prices than would be observed in the long run under perfect competition Since output was measured in value terms, this will lead to positive residuals Of 49 industries with positive residuals of more than 2 SE, we found 19 industries to be in the first group including wood engraving, gold and silver reducing, lapidary work, music publishing, glucose, starch, linseed oil, patent medicines, tin plate, brass, and lead. At least six were in the group characterized by expanding demand, namely, cordials and flavoring syrups (1909, 1914, 1919), oleomargarine (1914), perfumery (1919), and washing machines (1919). There were also at least 12 industries which were characterized both by imperfect competition and by expanding demand These were airplanes (1914), chewing gum (1919), cigars and cigarettes (1919), fountain pens (1914), photographic supplies and equipment (1904, 1909, and 1914), cash registers (1889), smelting and refining copper (1899, 1904), typewriters and supplies (1889) All these taken together accounted for 37 of the 49 high positive residuals. We concluded, therefore, that a very large majority of the major positive residuals, so far as the United States was concerned, were from precisely those industries that one would expect on a priori grounds. This strengthened the case for the production function as a description of normal competitive relationships.

---

[3] The Bronfenbrenner-Douglas study first introduced an analysis of the deviations of the actual from the theoretical as measured by the standard errors of estimate

Because of limited time, we did not make as detailed an inquiry into the 38 major negative residuals We would expect them to be from industries characterized by overexpansion of the stock of labor or capital, by declining demand, or by the use of predominantly low-skilled labor. Fifteen were found to be industries where the supply of labor was overabundant Among these industries were 10 cases in the flax, hemp, linen, jute, and oakum group. Three more were connected with cotton, which tends to use low-skilled labor Two, hammocks (1889) and nets and seines (1914), were in allied industries while three more, charcoal, waste, and the canning of oysters, were highly disagreeable This made a total of at least 21 of the major negative residuals which were due to an oversupply of labor or the use of low-skilled labor. Several more were due to contracting demand Among these were grindstones, millstones, and hooks and eyes We regret that it was not possible to make as full an analysis of this class of cases because of lack of time caused by developments which I shall later narrate

But enough evidence was accumulated to indicate that most of the major deviations of the actual from the theoretical values of *P* were explainable by dynamic departures from "normal" which, whether for good or evil, caused the "productivity" to depart markedly from the formula as a description of "normal" conditions The case for the use of the function was, therefore, further strengthened Moreover, since the influences of imperfect competition, expanding demand, oversupplies of labor, and contracting demand caused so many of the large errors, it seemed probable that many of the lesser deviations were also caused by those same factors

Charts showing the distributions of the actual products from about the line of the theoretical relationship were prepared and shown for all 15 American and dominion studies, but because of restrictions of space and expense only three could be published in the address. With one or two exceptions the records of the total of 3,558 observations strongly corroborated both the formula and function

I presented all of these findings in my presidential address to the American Economics Association in Chicago in 1947. At the conclusion of my address, I invited critics to check my statistics and conclusions and offered to put the necessary data at their disposal. To the best of my knowledge this has not been taken advantage of. Certainly, the previous objection that the conclusions were based on an insufficient number of cases could no longer be advanced.

The Years after the 1947 Address

On the very night of my presidential address in Chicago, the Democratic State Committee was meeting in that city to select candidates for U.S senator and for governor. Just as I had finished dressing and was about to go down with my wife to the meeting, the telephone rang. It was my friend, Colonel Jack Arvey, who told me that I had just been nominated for senator and asked if I would accept Getting a hurried nod from my wife, I thanked him and the committee and pledged that I would do my best. Then, as we walked down to the meeting hall, I remarked that it might mean the end of my scientific studies, and I found myself repeating a line from Othello, "O farewell forever now, the tranquil life " It had not been tranquil, but it was to prove even less so

This is not the place to describe what followed except to say that during my 18 years in the Senate, and for two frustrating years as chairman of a presidential commission, I was unable to work on the production function Then I was put out of commission for another year by the Asiatic flu, pneumonia, a heart attack, and finally, a moderately severe stroke As I recovered, I picked up the subject once more and discovered that the old opponents were now relatively silent, that the Cobb-Douglas function was being widely used, and that a host of younger scholars led by my former student, Paul Samuelson, his colleague Solow, and Marc Nerlove, the son of my friend and former colleague, Samuel Nerlove, were all pushing forward into new and more sophisticated fields I first urged younger scholars to carry out empirical studies for which there was a plethora of evidence in Canada, Australia, and New Zealand Together with Senator Proxmire, I urged the American statistical authorities and the Congress to resume the collection of capital statistics which the economists and statisticians had persuaded them to drop in 1920–21 Seeing no visible signs that this advice was being followed, Gunn and I then resolved to do it ourselves So nearly a half century after the original Cobb-Douglas study, and almost a quarter of a century after my presidential address, we launched a fourth sortie into the production function

We chose seven more recent years in Australia. These were fiscal years 1956, 1957, 1964, 1965, 1966, 1967, and 1968.[4] About 160 industries were covered each year or 1,123 in all These, when combined with the earlier studies, made a total of 2,496 for the cross-section analysis of the British Commonwealth and, with the 2,185 American cross-section

---

[4] We also obtained U.S data for 1967 from Rey B Madoo, who had employed them in his forthcoming thesis ("Production, Efficiency, and Scale in U.S. Manufacturing [1967], an Inter-Intra-Industry Analysis," University of California, Berkeley) Using these data to estimate an interindustry production function for 50 industries, I found $k = .60$ and $j = .40$, with an R^2 of 95.

TABLE 4

PRODUCTION FUNCTION FOR AUSTRALIAN MANUFACTURING, 1956, 1957, 1964–68

Fiscal Year	*N*	*k*	SE of *k*	*j*	SE of *j*	*k* + *j*
1956	159	615	.03	.365	.02	.980
1957	159	.610	.03	381	02	.991
1964	163	.595	.03	396	.03	.991
1965	161	.576	03	.414	.03	.990
1966	161	562	03	434	.03	996
1967	160	575	.03	425	03	1.000
1968	160	.536	.03	.456	.03	.992

observations, a grand total of 4,681 The contention that our conclusions were based on an insufficient number of cases would hardly seem to be tenable after all this.

The values of *k* and *j* and their probable errors of estimate for the more recent years are shown in table 4 The results are a corroboration of what had been found before. The values of *k* + *j* tend roughly to equal unity, indicating a tendency toward true constant returns But once again such slight deviations from unity as exist are on the minus side. These deviations, however, are smaller than the standard errors of the estimated coefficients

Let us turn now to an examination of the residuals in these more recent years.[5] Those industries whose estimated product was more than 2 SE away from actual product remained fairly constant over the 7 years. Actual product fell short of estimated for arms and ammunitions producers in all 7 years and for explosives manufacturers in 5 of the 7 years. Actual product exceeded estimates for pharmaceuticals, tobacco, soap, candles, and gramophones in all 7 years Since all of these are industries that advertise heavily, it is likely that the omission of goodwill from the capital stock has contributed to the underestimates in these cases They also were probably characterized by imperfect competition

The relationship of *W*/*P* to *k* is very important from the standpoint of distribution theory. Table 5 sets forth these values for 7 years in the 1950s and 1960s. The relationship between *k* and *W*/*P* is quite strong For the 2 years of the fifties *W*/*P* was only from 1 to 3 percentage points below *k*, while in the sixties the gap ranged from 2 to 7 percentage points.

We should not let these minor differences obscure the fact that a substantial degree of equality between *k* and *W*/*P* has been attained. That is the central fact, and it both gives further corroboration to the production function and tends to show that the distribution of the

---

[5] If the residuals from the regressions were distributed according to the normal distribution, we should have observed roughly 68 percent of them which were equal or smaller in size than 1 SE. In general, the residuals from the work reported here were distributed such that 70–75 percent of them were within 1 SE of the estimate.

TABLE 5

RELATIONSHIP BETWEEN $\frac{W}{P}$ AND k FOR AUSTRALIAN MANUFACTURING, 1956, 1957, 1964–68

	Production Function for Total Production	
Fiscal Year	k	$\frac{W}{P}$
1956	615	.602
1957	610	.581
1964	595	.527
1965	576	530
1966	562	528
1967	575	517
1968	536	514

product closely conforms to what, in a largely competitive society, we would expect the marginal productivity of labor to produce Both productivity and distribution therefore, mutually reinforce each other

There would be a still closer degree of agreement between W/P and k (or $k/[k + j]$) if it had been possible to include the payments to all outside workers Instead, as has been pointed out, W does not include outside salesmen, stationary or traveling Nor does it cover carters in the employ of the manufacturing establishments, nor storemen who sell products at retail from the factories These men are necessary parts of the production process But they receive wages and salaries which have to be paid before interest and profits can be distributed They, therefore, should be added to W to give a truer picture of the share received by labor By how much their inclusion would raise W/P is of course unknown. But, in my opinion, it would not be by enough to eliminate the difference between k and W/P

What of the Future?

The results of this study lend further corroboration to the accuracy of the production function as a description of manufacturing production and as a determinant of the distribution of the product—which is a separate but allied subject

One might still object that different industries have different production functions, but the degree of confirmation received by the logarithmic formulation with constant returns to scale in interindustry studies is very striking indeed. A number of studies have been performed that use cross-sectional observations on individual industries They avoid the possible problem of using diverse industries to estimate a single function, but their results are also in accord with those reported here

The most ambitious of these is by Benjamin Klotz, who used data for 17 four-digit manufacturing industries in the United States for 1957 and 1963 (Klotz 1970). While his estimated coefficients vary considerably, and while most industries show decreasing returns to scale, Klotz notes that no industry conclusively rejects the hypothesis of constant returns to scale. These are the same conclusions that we have been reaching with other bodies of data for almost 50 years

A similar study using 1963 data from individual establishments in Norway yields results very much like Klotz's (Griliches and Ringstad 1971) Griliches and Ringstad estimate a coefficient for labor of .865 and one for capital of 199, when all their observations are pooled Thus, they find very slightly increasing returns to scale They conclude "that it is very hard to improve upon the simple Cobb-Douglas form."

A considerable body of independent work tends to corroborate the original Cobb-Douglas formula, but, more important, the approximate coincidence of the estimated coefficients with the actual shares received also strengthens the competitive theory of distribution and disproves the Marxian Many of the original objections have been answered. Some remain

Certainly the crude Marxian slogan, "Capital is dead labor which vampire-like lives by sucking living labor and lives the more, the more labor it sucks," can no longer be accepted Capital itself is instead productive, not exploitative It will add to production in a communistic as well as a welfare capitalistic society The ethical question of who should own the capital and the form which such ownership should take is still open and should be decided on its merits But some of the denunciatory appeals can and should be omitted The issues are ethical, economic, and engineering, and the proper prophets of a new order are not Marx or Lenin, but Robert Owen and the British Fabians

It was my hope that the present study of 7 additional years would be only the prelude to a study of nearly 60 years (1912–70) From this I hoped to get final values for Australia of both k and j and their relationship to a true W/P I could also then see more fully what changes, if any, were caused by the business cycle and whether there were long-term changes in k, j, and W/P, as well as alteration in the pattern of residuals But the ravages of time raise serious doubts whether I can accomplish this I, therefore, still appeal to the younger generation of economists and statisticians for help. A quarter of a century ago there would have been distinct professional risks in such a venture The opponents of the production function were eminent, powerful, and determined. No such danger exists today The times favor such studies. I hope they may be made Despite all the difficulties, I intend to push on and complete, if I can, a long-time study of the function in Australia for the years 1912–70

References

Bronfenbrenner, M., and Douglas, P. H. "Cross Section Studies in the Cobb-Douglas Function 1909." *J P.E.* 47 (December 1939)· 761–85.

Browne, G. W. G. "The Production Function for South African Manufacturing Industry." *South African J. Econ.* 11 (1943)· 259.

Cobb, C. W. "Production in Massachusetts Manufacturing, 1890–1928." *J P E.* 38, no. 6 (December 1930) 705–7

Cobb, C. W., and Douglas, P H. "A Theory of Production " *A E R* 8, no 1, suppl (March 1928) 139–65

Daly, P., and Douglas, P. H. "The Production Function for Canadian Manufacturing Industry." *J. American Statis Assoc* 38 (1943) 78–86

Daly, P., Olson, E , and Douglas, P H. "The Production Function for Manufacturing in the United States in 1904 " *J P E* 51, no 1 (February 1943). 61–65.

Durand, D "Some Thoughts on Marginal Productivity with Special Reference to Professor Douglas " *J.P E* 45, no 6 (December 1937) 740–58

Griliches, Z , and Ringstad, V *Economies of Scale and the Form of the Production Function* Amsterdam North-Holland, 1971

Gunn, G T , and Douglas, P H "The Production Function for American Manufacturing in 1919 " *A E R.* 31 (March 1941) 67–80

——— "The Production Function for American Manufacturing for 1914." *J P.E.* 50, no 4 (August 1942) 595–602

Handsaker, M L , and Douglas, P H "The Theory of Marginal Productivity Tested by Data for Manufacturing in Victoria " *Q J E* 52 (November 1937 and March 1938). 1–36, 215–54

Klotz, B P. "Productivity Analysis in Manufacturing Plants " Bureau of Labor Statistics Staff Paper no 3, U.S Dept Labor, 1970

Leser, C. E. V "Production Functions and British Coal Mining " *Econometrica* 23 (October 1955) 442–46

Lomax, K S "Production Functions for Manufacturing Industry in the United Kingdom " *A E R* 40 (June 1950) 397–99

Mitchell, W C (ed), King, W I , Macauley, F R , and Knauth, O. W. *Income in the United States Its Amount and Distribution II, Detailed Report* New York Nat Bur Econ Res , 1922

Williams, J W , and Douglas, P H "Production Functions " *Econ Rec* 21 (1945)· 55–63

A Statistical Theory of Expenditures in Capital Maintenance and Repair

---

George C. Bitros

Bank of Greece

This study introduces an econometric model to explain the determinants of expenditures in capital maintenance and repair and, indirectly, to question the assumption implicit in most investment studies that such expenditures do not "matter" in the process of capital accumulation. The model is estimated by a consistent-systems technique, with data pertaining to the rolling stock of class-I line-haul railways in the United States from 1944–70 The empirical results indicate that maintenance expenditures are determined by gross additions to and retirements from the rolling stock as well as by the cost of funds and the rate of utilization. Moreover, in light of the uncovered trade-offs between maintenance expenditures and gross investment, the paper concludes that in estimating investment models, a proxy for maintenance expenditures should be included among the independent variables.

I. Introduction

In the aftermath of the Keynesian revolution economists made a tremendous effort to explain the behavior of almost every element making up the income-expenditure identity Studying the determinants of investment, in particular, became one of the two major focal points of research, the other being consumption Amidst all this activity, and given the possibility that a certain trade-off might exist among the various decisions relating to potential capital services, one would expect investment in new

This paper was written while I was associated with the National Bureau of Economic Research and New York University I am greatly indebted to the referee for his extremely valuable comments and criticisms Moreover, I should like to extend my gratitude to M Darby, S Fabricant, W Lawrence, and P Meller for giving me the benefit of their advice and to Mr. K. Hurdle of the Association of American Railroads, who provided me with some useful unpublished information Any errors of fact or interpretation are my own

[*Journal of Political Economy* 1976, vol 84, no 5]
© 1976 by The University of Chicago. All rights reserved

durables to be investigated in conjunction with expenditure on maintenance and repair. However, this has not been the case. Contemporary econometric analyses of investment, such as those by Meyer and Kuh (1957), Jorgenson (1963, 1965), Hickman (1965), and Eisner and Nadiri (1968), proceed as if the decision to alter the useful life of existing capital stock through maintenance and repair did not affect the decision to invest in new capital goods.[1] As a result, the cyclical behavior of maintenance expenditure along with its possible repercussions on investment and national income have been grossly ignored

The suppression of the maintenance and repair aspect of investment analysis appears to be based on two grounds First, there are the indications offered by Jorgenson (1963), Jorgenson and Stephenson (1967*a*, 1967*b*), and others that replacement investment is a constant proportion of the existing capital stock Obviously, if the combined effect of deterioration and maintenance and repair results in an evaporative death of capital at a constant rate, the influence of maintenance and repair can be parameterized, thus leaving investment as the sole factor of decision in relation to capital services The second is associated with the fact that data on expenditures for maintenance and repair are very hard to obtain Now, the investigations by Feldstein and Foot (1971), Eisner (1972), and Bitros and Kelejian (1974) have shown that replacement is not a constant proportion of existing capital stock [2] On the other hand, the study by Jorgenson and Handel (1971) reveals that, even when applied to sectors where relatively good maintenance and repair data are available, modern investment theory assumes the maintenance decision away It follows, therefore, that an attempt to reconsider the theory and to extend the evidence in this area is in order

In this connection, this paper constructs a model to explore the determinants of maintenance expenditures and, indirectly, to question the tacit assumption of investment studies that the expenditures on maintenance and repairs are irrelevant to the process of determining potential capital services. More specifically, the hypothesis considered here is that, given such forces as the utilization rate and the cost of funds, maintenance and repair expenditures, gross investment, and scrappage are simultaneously determined To test this hypothesis, the model is then estimated using data pertaining to the locomotives and freight cars of class-I railways in the United States from 1944 to 1970.

The empirical results obtained are as expected Maintenance expenditures are found to be significantly related to gross investment and scrappage as well as other cyclically sensitive variables. Moreover, comparison of the two-stage least-squares (2SLS) estimates of the basic equations with those obtained by applying ordinary-least-squares (OLS) reveals that the trade-offs which exist among maintenance expenditures, gross investment, and scrappage must be substantial. Hence, when estimating investment and scrappage equations, it is advisable that maintenance expenditures be included among the independent variables and, in addition, a simultaneous-equations estimating technique be used, since the "system" bias is likely to be considerable.

---

[1] An exception to this may have been Johansson (1961), who, according to the account given by Näslund (1966), has dealt with the relationship between the repair and service life policy and the investment decision. Parks (1974) has also dealt recently with an aspect of the problem in reference to the management of consumer durables

[2] Convincing theoretical arguments against the proportionality hypothesis have been provided by Feldstein and Rothschild (1974)

Section II presents the arguments on which the model specifications and coefficient sign expectations are based, Section III describes the nature of the data and defines the variables, and Section IV specifies the equations to be estimated and presents and interprets the empirical results. The conclusions are summarized in Section V

II. Theoretical Foundations of the Model

Maintenance, used either in the preventive or corrective sense of the term,[3] is the deliberate employment of resources in the form of labor, spare parts, and other materials to preserve the operative state of capital goods [4] As such it is a form of investment which entails certain costs and in return gives rise to a stream of future benefits How much is spent on the current upkeep of existing capacity should depend positively on the expected rate of return of maintenance resources vis-à-vis the return that might be obtained if these resources were employed to secure capital services through alternative courses of action By implication, the behavior of maintenance expenditures can be analyzed by focusing on the directional influences of forces which determine its relative profitability

Following Bitros and Kelejian (1974), who have considered the decisions relating to capital services, one hypothesis is that the relative profitability of maintenance depends on what the respective resources could earn if invested in new capital and on the economies that might be involved in scrapping. To reflect first on the effects of new capital goods, consider the appearance in a given production line of a new machine, which, because it incorporates the latest technological developments, has lower operating costs and/or produces higher quality products than the machine already in place As the analyses by Terborgh (1949) and Solow (1970) suggest, the ensuing competition between the two machines will cause the quasi rents of the old machine to decline Consequently, since the expenditure on maintenance is positively related to its rate of return and this is, in turn, negatively related to capital embodied technological change, gross investment, which is the carrier of such progress, should be negatively related to maintenance expenditures [5]

---

[3] The various types of maintenance and repair policies that can be followed are very well described in Nesvera (1970) For some additional works in this area, see Crane and Brown (1954), Eilon, King, and Hutchinson (1966), Brems (1968), and Ullman (1970)

[4] Note that according to this definition maintenance includes expenditures for regular upkeep and repairs of capital but does not include money spent for the purpose of neutralizing the effects of economic obsolescence Expenditures for this purpose would be included in modernization investment Where the line is drawn in practice between "regular" maintenance and "modernization" is arbitrary, as Fabricant points out (1938) Maintenance expenditures for either purpose may be indistinguishably charged to current expenses, to deferred charges, to depreciation reserves, or to the fixed-asset accounts Unless otherwise stated, the discussions below concern the "maintenance and repair" expenditures which, according to the accounting convention, are charged to current expenses or to deferred charges

Gross investment, however, may also be a proxy for demand conditions. Indeed, while, according to the evidence provided by Nadiri and Rosen (1973), the utilization rate acts as a buffer in the short run, in the long run permanent changes in demand lead to capacity adjustments Thus, given that the production and installation of capital goods takes time, if an increase in capacity is called for, the return on maintenance (and by implication maintenance expenditures) might rise pari passu with gross investment Similarly, where capacity must be curtailed, the expenditures on maintenance and the purchase of new durables might decrease simultaneously, since the profitability on both investment activities moves in the same direction Therefore, to the extent that gross investment is a proxy for demand effects, its negative influence on maintenance expenditures (suggested in the preceding paragraph) might be weakened or even be reversed altogether

Turning next to scrappage, its relationship to maintenance expenditure is expected to be a negative one, due to the way in which the profitability of maintenance depends on the life span of capital goods Specifically, from the models presented by Naslund (1966), Jorgenson, McCall, and Radner (1967), and Thompson (1968) it can be inferred that the faster capital goods are discarded from production the lower the rate of return on their maintenance becomes, since their service lives decline Hence, if scrappage changes, due to changes in the rate of technological obsolescence (and/or output demand), the profitability of maintenance varies in the opposite direction, and the amount of resources allocated to it should relate to scrappage negatively [6]

Now, before the analysis proceeds to consider other determinants of maintenance expenditures, a note is in order. From the arguments above it is clear that the influences of gross investment and scrappage on maintenance expenditures are not unidirectional. Rather, the former variables affect and are themselves affected by the latter This implies that, in the context of deciding upon the most economical manner in which to provide a desired level of capital services, the rates of maintenance expenditures, gross investment, and scrappage are complementary To allow, then, for possible trade-offs between maintenance expenditures on the one hand and gross investment and scrappage on the other, the latter variables will be treated as endogenous

---

[5] This relationship, however, may be weakened inasmuch as technological progress can be embodied For, if an old machine could be economically made as modern as a new machine and maintenance substituted for modernization, gross investment and maintenance expenditures would have a tendency to be related positively.

[6] In principle, maintenance expenditures could also be positively associated with scrappage This could occur if gross investment were a substitute for existing equipment while, at the same time, associated with rising output demand and utilization. Clearly maintenance expenditures per unit of surviving equipment would normally be highe than the maintenance requirements of new equipment installed It follows that increase

Maintenance expenditures may also be influenced by certain exogenous variables. One that comes immediately to mind is the cost of funds This variable, as measured by the interest rate, may affect maintenance expenditures either directly, through changes in their expected profitability, or indirectly, by altering the durability and hence the maintenance requirements of capital stock Since the present value of the benefits expected to accrue from maintenance decreases or increases as the cost of funds goes up or down, maintenance expenditures and the interest rate are negatively associated At the same time, however, as more durable goods are normally more expensive than the less durable ones, a drop in the cost of funds might cause a shift away from shorter-lived to longer-lived capital expenditures, and vice versa In turn, such a shift would imply a positive relationship between maintenance expenditures and the interest rate, since capital durability and maintenance requirements can be presumed to be related inversely There is no way of predicting with certainty what the net effect of these two influences will be Nevertheless, granted that the positive effect must operate slowly over time, due essentially to the difficulty involved in changing the age composition of capital stock in the short run, the negative effect will probably be the dominant one The expectation, then, is that, ceteris paribus, maintenance expenditures will be negatively related to the interest rate

Another such determinant is the intensity with which capital goods are used Presumably, since utilization depletes the output capacity of capital,[7] the more intense the usage is the more valuable the capital

---

maintenance expenditures, due to higher demand and utilization, could possibly go hand in hand with increased scrappage, due to the introduction of new equipment Although such a case is conceivable, I have considered it extremely doubtful For this reason, I have elected not to weaken the view taken in the text that maintenance expenditures and scrappage would be always negatively associated

[7] Marris (1964), Hirschman (1967), Winston and McCoy (1974), and others have advanced the view that the utilization of capital may be considered exogenous only in the short run, since in the long run it is probably an economic variable subject to the optimizing calculus of the firm Here I have decided to conceive of utilization as an exogenous force because I am interested in its short-term impact on maintenance expenditures For an example where utilization is treated as an endogenous variable, see Bitros and Kelejian (1974).

services and therefore, ceteris paribus, the higher the profitability of maintenance must be. On these grounds, then, the expectation is that the utilization rate and maintenance expenditures will be positively related As for the negative relationship between these two variables, which is suggested by the consideration that maintenance takes place predominantly when capital is idle, the likelihood is that its influence will be extremely weak The reason for this is that, as pointed out in Winston's (1974) review article, capital goods remain idle most of the time

In summary, the theoretical arguments presented in this section suggest that maintenance expenditures depend negatively on scrappage and the cost of funds and positively on the utilization rate With regard to gross investment, its relationship to maintenance expenditures can go either way because this variable stands for more than one opposite effect.

III. Nature of the Data and Definitions

The data used in the estimations were time series pertaining separately to the stock of locomotives and freight cars of class-I line-haul railways in the United States from 1944 to 1970 They were primarily obtained from the *Statistics of Railways in the United States*, published annually by the U S Interstate Commerce Commission Various issues of the *Transportation Manual*, released by Moody's Investors Service, Inc , served as a secondary source of information From this I extracted a few series which, though traceable to other U S Interstate Commerce Commission publications, were readily available there Finally, some minor discontinuities in a variable were filled in and an important series was put together with the help of unpublished information given to me by Mr K H Hurdle of the Association of American Railroads Relatively speaking (this area of research is characterized by a scarcity of pertinent data), this sample is most suitable for the test under consideration here

This last statement is based on at least three desirable features of the data The first is the consistency with which capacity is measured Given that the capacity of locomotives is rated in the physical and homogeneous unit of "tractive effort" [8] and that output is expressed in terms of ton miles carried—aside from the question of reliability of maintenance (for which the data do not account)—there must be no aggregation biases involved, so that the series of locomotive capacity is internally consistent over time The same is true of the capacity of cars, which is given in short tons[9] of freight-carrying ability The second attribute is that the capacity of locomotives and the capacity of freight cars are observed independently of the new additions to them, hence, the corresponding series of available capacity in locomotives and freight cars are also consistent with the intertemporal paths of the respective series of gross investment and scrappage for each of these two capital goods. Last, the data are characterized by an admirable continuity of defintions, both in the course of collection from individual companies and their aggregate presentation. For example, while the definitions originally accepted by this industry's uniform system of accounts[10] as to the delineation of maintenance and investment may have been somewhat arbitrary, one thing is certain: the definitions for maintenance and gross investment (as for most of the variables) have not changed during the period of the sample, so that the respective accounts have retained their intertemporal comparability.

---

[8] General Motors Corporation, in its manual *Pull—the Tractive Effort of Diesel Locomotives* (n d), defines "tractive effort" as the turning force at the rims of the wheels, if conditions are such that the wheels cannot slip

[9] A short ton is equal to 2,000 pounds

Turning now to the description of the individual variables, the time series for the maintenance variable of locomotives is derived through application of the definition $M_t^l = m_t^l/w_t$, where m_t^l represents the absolute level of current expenditures incurred for maintenance and repairs of locomotives, and w_t is an index of labor costs in the maintenance departments of railway companies. Following the same definition but replacing m_t^l by the level of current expenditures for maintenance of freight cars, m_t^c, gives the maintenance variable of freight cars, M_t^c. I would have liked to be able to deflate maintenance expenditures by an index involving the costs of spare parts, lubricants, anticorrosive materials, etc., as well as the cost of labor. Such an index, however, cannot be constructed because the prices of the former items are not available.

With reference to the capacity variables, that is, one for locomotives and another one for freight cars, two minor data imperfections had to be faced: (*a*) the data detailing tractive effort of owned and leased locomotives and tons of carrying ability of owned and leased freight cars were not reported separately, and (*b*) beginning with 1968, the reporting unit of locomotive capacity changed from "tractive effort" to "horsepower." In deriving the capacity of owned locomotives, the average tractive effort of leased locomotives was assumed to be the same as that of the locomotives in the total stock. Then, multiplying the number of leased locomotives by the average tractive effort of all locomotives and subtracting the product from the total tractive effort at the close of the year results in the variable of owned locomotive capacity, K_t^l. The capacity of owned freight cars, K_t^c, was also segregated from the total in a similar manner. However, had total capacity been used instead of owned capacity alone, the estimations would not have been materially affected, since the leased capacities of locomotives and freight cars over the same period never exceeded 18 percent and 19 percent of the total capacities, respectively[11] On the other hand, the observations on locomotive capacity for 1968, 1969, and 1970 were converted from horsepower to tractive effort by means of formula (1)[12]

---

[10] As every other regulated industry in this country, railway companies are required to follow certain rigidly defined accounting procedures in handling the various types of revenues and expenditures. The laws that describe these procedures are detailed in the U.S. Interstate Commerce Commission's *Uniform System of Accounts for Railroad Companies* (1957).

$$\text{Tractive effort (in pounds)} = \frac{\text{engine horsepower} \times 308}{\text{miles per hour}}, \qquad (1)$$

assuming that the locomotives purchased after 1967 could run as many miles per hour (mph) as the average locomotive added to the stock during the three preceding years[13]

Gross additions to locomotive capacity, I_t^l, consist of the tractive effort added via new units purchased or built plus units rebuilt and written into property accounts. Gross additions of freight car capacity, I_t^c, are made up of short tons of cargo-carrying capability added through purchases of new cars and cars built or rebuilt during the year Moreover, as to the scrapped capacity of locomotives, S_t^l, and freight cars, S_t^c, these variables consist of the tractive effort and tons of carrying ability, respectively, which are permanently withdrawn from service annually For completeness, table 1 gives the expenditures for maintenance as well as the capacity series on which the estimates are based

The series on utilization rates were by far the most difficult to derive Referring first to the utilization rate of locomotives, U_t^l, this was approximated by the ratio of gross ton miles carried to capacity Capacity, however, as indicated above, is measured in terms of tractive effort, so that it had to be converted into ton miles To accomplish this, I initially employed formula (1) in conjunction with the average speed at which freight and passenger trains run to convert the available tractive effort into capacity expressed in horsepower[14] Then, making use of the fact that one horsepower produces 1/0.1875 ton miles per hour,[15] I computed U_t as

---

[11] In addition, substitution of the total for the owned capacities in the estimations reported in this paper left the results essentially unaffected

[12] This formula was brought to my attention by Mr. K H Hurdle of the Association of American Railroads It is contained in the manual of General Motors Corporation and relates only to the tractive effort of diesel locomotives However, no bias results from applying it to the total tractive effort for 1968, 1969, and 1970 because dieselization of railways was almost 100 percent complete by 1963

[13] More specifically, the operation involved three steps First, I calculated the miles per hour which would consistently convert the reported horsepower of gross additions to locomotive capacity during 1965, 1966, and 1967 into the declared capacity in terms of tractive effort Second, using the average of miles per hour as estimated in the first step for 1965, 1966, and 1967 and the reported horsepower of gross additions to locomotive capacity for 1968, 1969, and 1970, I derived the gross tractive effort during these years Finally, multiplying the locomotives retired within each year by the average tractive effort of all locomotives at the close of the previous year and subtracting the consecutive products from the gross tractive effort for 1968, 1969, and 1970 as estimated above resulted in the net changes of locomotive capacity during these 3 years

[14] The conversion of tractive effort to ton miles may have been somewhat unstable over time for two reasons The first relates to the hours during which locomotives can

$$U_t^l = \frac{\mathrm{GTM}_t}{(\mathrm{HP}_t/0.1875) \times (1 - \mathrm{USL}_t) \times 8{,}760}, \tag{2}$$

where

GTM_t = gross ton miles at time t (gross in the sense that in addition to the useful load it includes the weight of locomotive and tenders as well as the weight for both passenger and freight train cars),
HP_t = capacity measured in horsepower at time t,
USL_t = ratio of unserviceable to serviceable locomotives at t, and
8,760 = the total number of hours contained in a year

The utilization rate of freight cars was computed by means of the following formula

$$U_t^c = \frac{\mathrm{NTM}_t^c}{K_t^c \times \mathrm{AFTS}_t \times (1 - \mathrm{USC}_t) \times 8{,}760}, \tag{3}$$

where

NTM_t^c = net ton miles of freight cars at t (net in the sense that it includes only the useful load),
AFTS_t = average freight train speed at t,
USC_t = ratio of unserviceable to serviceable freight cars at t, and
K_t^c = capacity of freight cars in short tons at t

The assumption that serviceable locomotives and freight train cars can

---

theoretically roll. Clearly, if the potential time during which locomotives could roll annually is not a fixed number of hours (as hypothesized here) but varies from year to year, the relationship between ton miles per hour and available annual ton miles of capacity would be shifting. Second, formula (1) is based on an efficiency parameter for diesel locomotives, which has the value 308. The value of this parameter would be affected by any changes in the relationship between tractive effort and horsepower. For example, to the extent that steam locomotives may have been less efficient in transforming horsepower to tractive effort than diesel locomotives, or to the extent that technological change may have improved the efficiency of diesel locomotives, the value of this parameter would change accordingly.

[15] To see why this is so, combine the definitions that (*a*) 1 foot = 1/5,280 mile, (*b*) 1 second = 1/3,600 hour, and (*c*) 1 pound = 1/2,000 ton with (*d*) 1 hp = 550 (foot × pounds)/seconds, as given in, say, Shortley and Williams (1965). Substituting (*a*), (*b*), and (*c*) into (*d*) gives

$$1\ \text{hp} = \left(550\ \frac{\text{ton}}{2{,}000}\ \frac{\text{mile}}{5{,}280}\right) \Big/ \frac{\text{hour}}{3{,}600} = 0.1875\ \frac{\text{ton miles}}{\text{hour}}$$

I am indebted to Mr. D. Kokkinos for having double-checked these operations on my behalf.

TABLE 1

MAINTENANCE EXPENDITURES AND LOCOMOTIVE AND FREIGHT CAR CAPACITY DATA PERTAINING TO THE CLASS I RAILWAYS IN THE UNITED STATES, 1944–70

YEAR	LOCOMOTIVES					FREIGHT CARS				
	Maintenance Expenditures ($ millions)	Tractive Effort (pounds)				Maintenance Expenditures ($ millions)	Freight-carrying Capacity (short tons)			
		Owned	Leased from Others	Newly Added	Retired		Owned	Leased from Others	Newly Added	Retired
1944	585	2,162,336	155,165	86,638	23,549	369	84,474	5,486	2,121	1,131
1945	577	2,176,486	152,996	66,531	52,381	365	84,500	5,372	2,063	3,038
1946	576	2,158,196	156,815	45,480	63,770	373	83,753	5,638	1,993	2,740
1947	604	2,135,187	153,377	89,103	112,112	399	83,731	5,493	3,041	3,063
1948	631	2,172,033	148,841	141,872	105,026	456	86,076	5,218	5,187	2,842
1949	542	2,143,583	154,760	191,030	219,480	428	85,831	6,130	4,206	4,451
1950	563	2,166,862	151,486	185,158	161,879	484	84,543	5,922	1,902	3,190
1951	635	2,193,986	142,948	207,053	183,774	591	86,635	6,036	4,444	2,352
1952	599	2,094,966	123,193	190,949	289,969	586	88,171	5,372	4,713	2,177
1953	563	2,028,478	75,981	127,006	193,494	631	90,143	5,027	3,790	1,818
1954	476	1,935,696	60,029	73,041	165,823	497	88,596	4,600	1,619	3,166
1955	481	1,912,673	97,942	77,859	100,882	546	84,897	6,334	2,019	5,718
1956	522	1,779,632	95,957	81,001	214,042	572	86,011	6,150	3,660	2,546
1957	513	1,754,707	107,685	65,931	94,856	569	88,239	6,844	5,186	3,958
1958	454	1,698,065	113,820	34,588	91,230	467	87,671	6,815	2,353	2,921
1959	471	1,706,757	106,132	61,236	52,544	516	84,932	7,318	2,016	4,755
1960	462	1,685,633	102,698	31,672	52,796	488	83,679	8,267	2,323	3,576
1961	430	1,701,462	90,482	16,942	1,113	438	81,174	8,118	1,868	4,373
1962	445	1,667,062	97,730	40,772	75,172	451	79,491	7,725	2,341	4,024
1963	448	1,650,634	105,849	46,604	63,032	442	77,440	8,503	3,038	5,089
1964	442	1,629,614	138,038	76,960	97,980	447	76,417	10,254	4,655	5,678
1965	444	1,604,917	154,385	71,984	96,681	438	77,554	10,768	4,997	3,860
1966	457	1,666,040	151,396	102,027	40,904	459	80,614	10,827	6,180	3,120
1967	455	1,643,785	167,399	70,994	93,249	448	81,776	11,892	5,508	4,346
1968	456	1,599,638*	201,155*	39,379*	83,526*	493	80,797	12,752	3,347	4,326
1969	488	1,584,921*	229,482*	54,484*	39,767*	534	80,623	13,723	3,723	3,897
1970	541	1,538,389*	280,653*	45,906*	92,438*	584	80,298	15,257	5,945	5,819

SOURCE.—U.S. Interstate Commerce Commission, *Statistics of Railways in the United States* 1944–1970

* To obtain these figures of tractive effort I converted horsepower ratings of locomotives by means of formula (1) and followed the steps described in n. 13

operate annually for 8,760 hours is an exaggeration; the figure was used because actual data on the hours rolling stock was available for service could not be obtained.[16]

Six other variables were also employed: (1) the interest rate, r_t, measured by Moody's composite yield on railroad bonds, (2) the average age of locomotives, A^l_t, and that of freight cars, A^c_t, approximated by the ratios of their net to gross fixed asset values, respectively, (3) the taxation rate, u_t, obtained by dividing federal, state, and local income tax accruals by operating income before taxes, (4) the variable of gross funds, F_t, arrived at by adding profits after taxes plus capital consumption allowances; (5) the relative cost of leasing versus owning equipment, L_t, represented as a ratio of rentals paid over the price of investment per unit of capacity; and (6) the percentage of total ton miles carried by trucks in the United States, MVT_t, employed to approximate the competitive pressure on the railway industry by other transportation modes

IV. Empirical Results

The statistical model to be estimated is

$$\frac{M_t}{K_{t-1}} = a_0 + a_1 \frac{I_t}{K_{t-1}} + a_2 \frac{S_t}{K_{t-1}} + a_3 U_t + a_4 r_t + \gamma X_t + v_t, \quad (4)$$

where, abstracting from the superscripts of variables for convenience, M_t stands for maintenance expenditures, K_{t-1} denotes output capacity, I_t and S_t are gross additions to and scrappage from capacity, respectively, U_t is the utilization rate, r_t represents the interest rate, X_t is a vector of additional variables, and v_t is a disturbance term Whether the model is fitted to the data for locomotives (i e , variables with the superscript l) or to those for freight cars (i e , variables with the superscript c), my expectations concerning the coefficients in (4) are $a_2 < 0$, $a_3 > 0$, and $a_4 < 0$ However, I have no expectations with respect to coefficient a_1.

Table 2 shows the 2SLS estimates obtained by fitting equation (4) to

---

[16] The theoretical capacity of locomotives and freight cars is difficult to establish anyway According to Petersen (1970), for example, the capacity of freight cars depends, among other things, on the following (1) total number of freight cars, (2) freight-car capacity and volume dimensions, (3) train speed, ruling grades, and location of passing sidings, (4) number of hours during which cars are available to the railroads, (5) freight-car routings, (6) carload mix which affects the type of cars used, the number of empty cars back hauled, and the switching and interchange facilities used, (7) train scheduling and make-up; (8) layout of classification yards, the location and length of receiving and departure tracks, condition of the rails, ties, and ballast, the efficiency of yard communication systems, and the availability of alternate tracks. Most of these factors cannot be measured so that for empirical research certain compromises are invariably adopted. For two alternative approaches to defining train capacity, see Barger (1948) and U.S. Interstate Commerce Commission (1952)

TABLE 2

2SLS Estimates of the Maintenance Equation for Locomotives

Equations	Explanatory Variables								R^2	S.E	D-W
		Endogenous		Exogenous and Predetermined							
	Constant	I_t^l/K_{t-1}^l	S_t^l/K_{t-1}^l	r_t	U_{t-1}^l	$(I_t^l/K_{t-1}^l)_{-1}$	$(S_t^l/K_{t-1}^l)_{-1}$	$(M_t^l/K_{t-1}^l)_{-1}$			
4.2.1	−4.81	72.63	−75.90	−176.55	67.10				.69	2.03	1.73
	(−0.87)	(1.85)	(−2.16)	(−5.37)	(4.15)						
4.2.2	−6.69	93.67	−21.94	−161.72	68.55	−74.02			.80	1.66	1.85
	(−1.45)	(2.77)	(−0.55)	(−5.78)	(5.16)	(−1.97)					
4.2.3	−2.60	70.28	−39.32	−177.89	61.20		−38.76		.76	1.84	1.37
	(−0.49)	(1.98)	(−0.91)	(−5.97)	(3.98)		(−1.26)				
4.2.4	−3.79	84.48	−93.16	−164.22	61.00			0.10	.71	2.29	1.91
	(−0.64)	(1.82)	(−2.22)	(−4.31)	(3.83)			(0.75)			
4.2.5	−2.62	49.24		−175.41	60.14		−57.64		.72	1.94	2.12
	(−0.46)	(1.72)		(−5.59)	(3.71)		(−2.40)				

NOTE.—Aside from the lagged values of maintenance, gross investment, and rate of utilization variables, six additional instruments were used in the estimation. These were RR_t, rate of return to railroad capital, MVT_t, percentage of ton miles carried by motor vehicles, USL_t, percentage of unserviceable locomotives, $(q_r^l/q_o^l)_t$, cost of leasing over cost of owning locomotive capacity, and the ratio of output price over the cost owning locomotive capacity $(P/q_o^l)_t$. The figures in the parentheses are t-ratios, S.E. denotes the standard error of the regression, and D-W stands for the Durbin-Watson statistic. For significance at the one-tail 5 percent level, the t-ratio for 22 df should be at least 1,717. Upon conversion of tractive effort from pounds to short tons the figures in table 1 would indicate that the ratio of locomotive to freight car capacities during the period under consideration was roughly 1:100. Hence, given that railroads spent approximately similar amounts for the annual maintenance of each of these two assets, had I measured maintenance expenditures for locomotives in simple dollars the parameter estimates shown in this table would have been generally in the order of 100 times the corresponding estimates for the maintenance expenditures of freight cars given in table 3. Consequently, to avoid scaling inconsistencies I have measured the maintenance expenditures of locomotives in hundreds of dollars. The capacity variables of both locomotives and freight cars are measured in short tons.

maintenance expenditures for locomotives.[17] It can be seen that these results are consistent with both the sign expectations outlined above and the hypothesis that maintenance expenditures, gross investment, and scrappage are simultaneously determined Specifically, the coefficients of scrappage and interest rate variables are uniformly negative throughout and that of the utilization rate is consistently positive At the same time, all variables in equation (4 2 1) are statistically significant at the 5 percent level, with the proviso that a one-tail test would be appropriate for the variable of gross investment

These results also reveal that none of the alternative lag structures considered improved the specification of model (4 2 1) In this respect, the lack of statistical significance of the coefficient for the lagged value of the dependent variable in eq (4 2 4) should be particularly noted. This suggests that, contrary to the preconception that maintenance-funds allocations per unit of capital would be characterized by inertia, in actuality there may not be any distributive lag structure relating the maintenance variable to its determinants Since, therefore, maintenance expenditures for locomotives are best explained by (4 2 1), the latter will be retained as the basic model

As could be expected, the statistical behavior of maintenance expenditures for freight cars is almost identical with that identified above for locomotives From table 3, which gives the 2SLS estimates of the model from data pertaining to freight cars, it may be observed that the results are again in general agreement with the theory Specifically, the coefficients of scrappage, interest rate, and utilization variables in equation (4 3 3) have all the postulated signs and are significant at the 5 percent one-tail test (with the exception of the coefficient for the utilization variable, which is significant even at the 5 percent two-tail test) Similarly, as in (4 2 1), the coefficient of gross investment is positive and statistically significant at the one-tail 5 percent level

The only difference between (4 2 1) and (4 3 3) lies in the presence of the lagged values of gross investment and scrappage variables in the second equation This finding might be interpreted as revealing the existence of some delayed impact of these forces on the maintenance expenditures of freight cars due, perhaps, to sharper underlying differ-

---

[17] As specified above, the maintenance model is conceived to be embedded in a system comprising a gross investment and a scrappage equation However, the implementation of the 2SLS procedure on (4) does not require the specification of the entire system of equations In particular, aside from the equation of interest, all one need specify is a set of instruments which are either contained in the system or are correlated with the predetermined variables of that system Thus, following Zellner (1957) and Black and Kelejian (1970), I have estimated (4) by specifying a number of such instruments These instruments are given at the bottom of table 2 for the maintenance equation of locomotives and in table 3 for the estimates referring to the maintenance expenditures of freight cars

TABLE 3

2SLS ESTIMATES OF THE MAINTENANCE EQUATION FOR FREIGHT CARS

EQUATIONS	EXPLANATORY VARIABLES								R^2	S.E.	D-W
		Endogenous		Exogenous and Predetermined							
	Constant	I^c_t/K^c_{t-1}	S^c_t/K^c_{t-1}	r_t	U^c_{t-1}	$(I^c_t/K^c_{t-1})_{-1}$	$(S^c_t/K^c_{t-1})_{-1}$	$(M^c_t/K^c_{t-1})_{-1}$			
4.3.1	7 10 (1.56)	7 34 (0 28)	−62 76 (−1.50)	−127 59 (−3 86)	92.32 (2.38)			.	77	1.57	1 62
4.3.2	9 61 (2.11)	37.25 (1 33)	−85 42 (−2.06)	−110 15 (−3 33)	80 47 (2 11)	−54 43 (−2 44)			.80	1.53	1.67
4 3.3	11 69 (2.70)	48 42 (1 83)	−72 17 (−1 85)	−103.05 (−3 35)	70 37 (1 98)	−58 61 (−2 83)	−50 41 (−1 98)		.83	1.41	1.72
4 3.4	8.41 (2 06)	69 87 (2 53)	−65 56 (−1 73)	−24 52 (−0 57)	−15 47 (−0 31)	−65 84 (−3 23)		64.78 (2.74)	.84	1.37	2.38
4.3.5	7 79 (1.75)	26 14 (0 99)	−12 35 (−0 28)	−146 07 (−4 31)	85 31 (2 27)		−49 88 (−1 82)		.80	1.52	1.56

NOTE.—*The set of instruments used to obtain these estimates was basically the same as the one described in the note to table 2 The differences were that to the subset of lagged* variables I added scrappage, and the variables USL_t, $(q^l_r/q^l_o)_t$ and $(P/q^l_o)_t$ were replaced by the percentage of unserviceable freight cars, USC_t, cost of leasing over cost of owning freight car capacity, $(q^c_r/q^c_o)_t$, and the ratio of output price over the cost of owning freight car capacity, $(P/q^c_o)_t$, respectively See note to table 2.

ences between locomotives and freight cars in the timing of short-term demand effects and long-term maintenance requirements. However, the lack of a general lag structure relating the maintenance-expenditures variable to its determinants is ascertained again, since the introduction of M^c_{t-1} into (4.3.3) worsens the specification For this reason I have preferred to retain (4.3 3) as the basic maintenance model for freight cars.

In concluding the comments on the statistical properties of the estimates presented above, one interesting aspect should be emphasized This is the tendency of a positive coefficient on current gross investment accompanying a negative coefficient on lagged gross investment as shown by (4.2.2) and (4 3 3). The positive coefficient suggests that current capital and maintenance expenditures are probably dominated by common demand factors The negative coefficient, on the other hand, indicates that equipment installed last year is likely to reduce the profitability of maintaining old equipment. Hence, in juxtaposition, the two signs might be taken as evidence in support of the speculation advanced in Section II that there is a separate demand and substitution effect in the relation between gross investment and maintenance expenditures

To shed some light on the economic significance of the findings, equations (4 2 1) and (4 3 3) were utilized to estimate the elasticities given in table 4 The salient feature here is the extraordinary responsiveness shown by maintenance expenditures for locomotives to changes in their utilization Specifically, a 10 percent increase in the average rate of utilization would be expected to increase maintenance expenditures by 26.23 percent per unit of tractive effort In terms of 1967 dollars this would mean, for example, an increase of $223 per short ton of tractive effort or an increase of more than $204 million in the average locomotive maintenance bill. Although smaller than the effect just discussed, the impact of the other variables on maintenance expenditures is also significant A 10 percent increase in the average scrappage rate of freight cars, for example, would induce a 5 98 percent decline in their maintenance expenditures, 3 59 percent of which would take place within the year of change, with the remainder realized in the following year. Translating these percentages into 1967 dollars, had such a change occurred in a given year of the sample period, by the following year it would have subtracted $ 54 from the maintenance of freight cars per short ton of carrying capacity, that is, it would have reduced the annual maintenance bill for this asset by approximately $45 million

Table 4 indicates, further, that maintenance expenditures may be quite sensitive to changes in the interest rate Note that it is difficult to tell what the true magnitude of the interest elasticity is because (as it will be pointed out later on) the coefficient of this variable contains a substantial amount of time trend, the same observation holds for the elasticity of utilization in (4.3.3), which is also trend dominated Despite these limitations, (4.2.1)

TABLE 4

ELASTICITIES OF MAINTENANCE EXPENDITURES PER UNIT OF CAPACITY ESTIMATED AT THE MEAN VALUE OF THE VARIABLES

NAME OF ASSET AND NUMBER OF EQUATION FROM TABLES 2 AND 3	INDEPENDENT VARIABLES					
	I_t/K_{t-1}	S_t/K_{t-1}	U_{t-1}	r_t	$(I_t/K_{t-1})_{-1}$	$(S_t/K_{t-1})_{-1}$
Locomotives (4.2 1)	380	−.513	2.623	− 928	.	
Freight cars (4 3 3)	.232	−.359	.851	− 492	−.265	− 239

and (4 3 3) have been found to be rich in economic implications. In addition, they confirm fully Hultgren's (1948) conclusions, according to which maintenance expenditures in this industry are essentially procyclical

The basic equations (4 2 1) and (4 3 3) were then subjected to two additional tests First, the list of regressors was expanded to include other variables which might represent either extra forces influencing the maintenance expenditures of locomotives and freight cars or different measures of forces already accounted for Among the variables considered in this connection were the rate of return to capital in the railroad industry, gross and net funds, the relative cost of leasing versus owning rolling stock, the average age of locomotives and freight cars, and the percentage of total ton miles carried by trucks in the United States From the empirical results obtained in the various runs of the basic equations in conjunction with these variables no obvious improvement was observed, and, therefore, the aforementioned equations were retained

However, one particular result of this series of experiments is worth mentioning When time was entered as an additional explanatory variable, the estimates obtained for the basic equations of the model were.

$$
\begin{aligned}
\frac{M_t^l}{K_{t-1}^l} = {} & \underset{(\ 83)}{3\ 79} + \underset{(1.91)}{46\ 00}\,\frac{I_t^l}{K_{t-1}^l} - \underset{(-2\ 64)}{58\ 26}\,\frac{S_t^l}{K_{t-1}^l} + \underset{(.13)}{7\ 78}\,r_t \\
& + \underset{(2\ 45)}{35\ 65}\,U_{t-1}^l - \underset{(-3\ 62)}{41}\ t
\end{aligned}
\tag{4.2.1'}
$$

$$R^2 = .87, \qquad \text{S E} = 1\ 36, \qquad \text{D-W} = 2.18$$

$$
\begin{aligned}
\frac{M_t^c}{K_{t-1}^c} = {} & \underset{(5\ 09)}{17\ 72} + \underset{(3\ 54)}{72.09}\,\frac{I_t^c}{K_{t-1}^c} \underset{(-1.88)}{-53\ 61}\,\frac{S_t^c}{K_{t-1}} \\
& + \underset{(1\ 39)}{60.29}\,r_t - \underset{(-\ 86)}{32.15}\,U_{t-1}^c - \underset{(-2.05)}{32\ 41}\left(\frac{I_t^c}{K_{t-1}^c}\right)_{-1} \\
& + \underset{(.59)}{1.26}\left(\frac{S_t^c}{K_{t-1}^c}\right)_{-1} - \underset{(-4.05)}{.48}\ t.
\end{aligned}
\tag{4.3.3'}
$$

$$R^2 = 92, \qquad \text{S E} = 1.01, \qquad \text{D-W} = 1\ 91$$

This indicates that the time variable substituted completely for the role of the interest rate in both equations. In addition, it eliminated the influence that the variables of utilization and lagged scrappage appeared to have in (4.3.3). Apparently, the large negative effect of the interest rate and, perhaps, that of utilization in the maintenance equation of freight cars reflect a considerable degree of time trend. Although (4.2.1) and (4.3.3) will be taken here to represent the best specification of the model, the signs and relative magnitudes of the trend-dominated variables will remain uncertain until more evidence is brought to bear on their structural relationship with maintenance expenditures.

In the second test, and in order to see how sensitive the results are to the estimation technique, I reestimated (4.2.1) and (4.3.3) by the OLS procedure. The results are given in (4.2.1″) and (4.3.3″) below,

$$\begin{aligned} \frac{M_t^l}{K_{t-1}^l} = &\underset{(-1.93)}{-7.43} + \underset{(0.59)}{10.43}\frac{I_t^l}{K_{t-1}^l} \underset{(-1.59)}{-21.57}\frac{S_t^l}{K_{t-1}^l} \\ &- \underset{(-7.62)}{177.32}\, r_t + \underset{(6.45)}{74.01}\, U_{t-1}^l \end{aligned} \tag{4.2.1″}$$

$$R^2 = 0.84, \qquad \text{S.E.} = 1.45, \qquad \text{D-W} = 1.42$$

$$\begin{aligned} \frac{M_t^c}{K_{t-1}^c} = &\underset{(2.11)}{7.52} + \underset{(.79)}{16.33}\frac{I_t^c}{K_{t-1}^c} - \underset{(-1.46)}{34.99}\frac{S_t^c}{K_{t-1}^c} - \underset{(-4.66)}{113.48}\, r_t \\ &+ \underset{(3.45)}{103.18} U_{t-1}^c - \underset{(-2.43)}{43.53}\left(\frac{I_t^c}{K_{t-1}^c}\right)_{-1} \\ &- \underset{(-2.12)}{47.90}\left(\frac{S_t^c}{K_{t-1}^c}\right)_{-1} \end{aligned} \tag{4.3.3″}$$

$$R^2 = .84, \qquad \text{S.E.} = 0.134, \qquad \text{D-W} = 1.74$$

where the figures in parentheses beneath the parameter estimates are t-ratios. In comparing the OLS estimates to those obtained by applying 2SLS, some very substantive differences should be noted. Specifically, if the basic equations had been estimated by the ordinary least-squares procedure I would have concluded that neither the gross investment nor scrappage variables were significant forces in determining maintenance expenditures. Now, however, the conclusion is that the systems "bias" in estimating maintenance, investment, and scrappage equations may be substantial, so that a systems-estimation technique should be used.

V. Summary and Conclusions

The objective of this paper has been to explore the determinants of expenditures in capital maintenance and repair. By first constructing a model and then using it to explain the behavior of class-I railways in keeping up their rolling stock, I have found that maintenance expenditures are determined to a large extent by gross additions to and retirements from the rolling stock as well as by the cost of funds and the rate of utilization. Moreover, the evidence has revealed that, for at least the rolling stock of this industry, the trade-offs that exist among maintenance expenditures, gross investment, and scrappage are substantial

The implications of these trade-offs concerning the assumption of investment studies that maintenance policies do not "matter" should be obvious. They indicate, although only indirectly, that if the interactions found among the variables referred to above exist for classes of capital assets *more general than rolling stock*, the traditional estimates of investment equations, where maintenance expenditures are parameterized, might involve serious biases However, to ascertain the existence and nature of these biases, gross investment equations will have to be estimated using maintenance expenditures as an independent variable and more aggregate data than those utilized here

Finally, I have argued that the method by which maintenance equations are estimated makes a significant difference In this respect, the conclusion is that system-estimating techniques are to be preferred This conclusion will also be supported in a forthcoming paper, where a system consisting of three equations corresponding to maintenance, gross investment, and scrappage is estimated by the three-stage least-squares (3SLS) procedure

References

Barger, H *The Transportation Industries 1889–1946* New York. Nat Bur. Econ. Res., 1948

Bitros, G C, and Kelejian, H H. "On the Variability of the Replacement Investment Capital Stock Ratio Some Evidence from Capital Scrappage." *Rev. Econ. and Statis* 56 (August 1974) 270–78

Black, S. W, and Kelejian, H H. "A Macro Model of the U.S Labor Market." *Econometrica* 38 (September 1970) 712–41.

Brems, H *Quantitative Economic Theory*. New York· Wiley, 1968

Capron, W. M, ed *Technological Change in Regulated Industries*. Washington. Brookings Inst, 1971.

Crane, R. R., and Brown, F. B. "Theory of Maintenance of Rolling Stock" *Mechanical Engineering* 77 (September 1954): 999–1000 and 1009.

Eilon, S.; King, G. R, and Hutchinson, D E. "A Study in Equipment Replacement." *Operational Res. Q.* 17 (March 1966) 59–71.

Eisner, R. "Components of Capital Expenditures. Replacement and Modernization versus Expansion." *Rev. Econ. and Statis.* 54 (August 1972). 297–305.

Eisner, R, and Nadiri, M. "Investment Behavior and Neo-classical Theory." *Rev. Econ. and Statis.* 50 (August 1968). 369–82.

Fabricant, S. *Capital Consumption and Adjustment* New York. Nat. Bur. Econ. Res., 1938.

Feldstein, M. S., and Foot, D. "The Other Half of Gross Investment. Replacement and Modernization Expenditures." *Rev. Econ. and Statis.* 53 (February 1971). 49–58.

Feldstein, M. S , and Rothschild, M "Towards an Economic Theory of Replacement Investment." *Econometrica* 42 (May 1974). 393–423.

General Motors Corporation *Pull—the Tractive Effort of Diesel Locomotives.* 7th ed. Lagrange, Ill General Motors, n d.

Hickman, B *Investment Demand and U S Economic Growth* Washington Brookings Inst., 1965.

Hirschman, A D. *Development Projects Observed* Washington. Brookings Inst , 1967.

Hultgren, T. *American Transportation in Prosperity and Depression.* New York Nat. Bur Econ Res , 1948

Johansson, S. E. *Skatt—Investering—Vardering, en Teoretisk Studie.* Stockholm Företagsekonomiska Forskningsinstitutet vid Handelshögskolan, 1961

Jorgenson, D. W. "Capital Theory and Investment Behavior " *A E R* 53 (May 1963)· 247–59

——— "Anticipations and Investment Behavior " In *The Brookings Quarterly Econometric Model of the United States,* edited by James S Duesenberry et al. Chicago Rand-McNally, 1965.

Jorgenson, D W , and Handel, S. S "Investment Behavior in U S. Regulated Industries " *Bell J Econ and Management Sci* 2 (Spring 1971) 213–64

Jorgenson, D W , McCall, T T , and Radner, R *Optimal Replacement Policy* Chicago Rand-McNally, 1967

Jorgenson, D W , and Stephenson, T A "Investment Behavior in U S Manufacturing 1947–1960 " *Econometrica* 35 (April 1967) 169–220 (*a*)

——— "The Time Structure of Investment Behavior in United States Manufacturing 1947–1960 " *Rev Econ and Statis* 49 (February 1967) 16–27 (*b*)

Marris, R *The Economics of Capital Utilization A Report on Multiple-Shift Work.* London Cambridge Univ Press, 1964

Meyer, J , and Kuh, E *The Investment Decision* Cambridge, Mass Harvard Univ. Press, 1957

Moody's Investors Service, Inc *Transportation Manual* New York Moody's, various

Nadiri, M I , and Rosen, S *A Disequilibrium Model of Demand for Factors of Production* New York Nat Bur Econ Res , 1973

Naslund, B. "Simultaneous Determination of Optimal Repair Policy and Service Life." *Swedish J Econ* 68 (June 1966) 63–73

Nesvera, V *A Study on Renewal, Repair, and Maintenance* New York United Nations Indus Development Org Document no ID/WG 62/22, 1970

Parks, R. W "Automobile Scrapping Rates The Choice between Maintenance and Built-in Durability." Discussion Paper no 74-12, Univ Washington, Inst Econ Res , 1974

Petersen, D L "Economic Effects of Technological Innovations on Railroads, 1947–1963 " Working Paper no 15, Case Western Reserve Univ , 1970.

Shortley, G , and Williams, D *Elements of Physics* Englewood Cliffs, N J. Prentice-Hall, 1965.

Solow, R. M. *Growth Theory An Exposition* New York Oxford Univ Press, 1970.

Terborgh, G *Dynamic Equipment Policy* New York McGraw-Hill, 1949.

Thompson, G L. "Optimal Maintenance Policy and Sale Date of a Machine." *Management Sci.* 14 (May 1968) 543–50

Ullman, A *Maintenance or Renewal* New York United Nations Indus Development Org , Document no. ID/WG 62/29, 1970

U.S. Interstate Commerce Commission. *Statistics of Railways in the United States.* Washington: Government Printing Office, various.

———. *The Capacity and Capital Requirements of the Railroad Industry.* Washington: Government Printing Office, 1952.

———. *Uniform System of Accounts for Railroad Companies, as Amended January 1, 1957.* Washington. Government Printing Office, 1957.

Winston, G. C. "The Theory of Capital Utilization and Idleness." *J. Econ. Literature* 12 (December 1974). 1301-20.

Winston, G C., and McCoy, T. D. "Investment and the Optimal Idleness of Capital." *Rev Econ Studies* 41 (July 1974) · 419–28.

Zellner, A. "The Short-Run Consumption Function." *Econometrica* 25 (October 1957): 552–67.

Temporary Layoffs in the Theory of Unemployment

Martin Feldstein

Harvard University

The typical worker who is laid off is soon rehired by his original employer. This important and generally unnoticed fact requires a major reevaluation of our current theories of unemployment. This paper develops a theory of temporary layoffs. Specific attention is given to the question of why employment is reduced instead of hours. The role of unemployment insurance and of taxes is examined in detail.

Most workers who are laid off are subsequently rehired by their original employers. This important and generally unnoticed fact requires a major reevaluation of our current theories of unemployment.[1]

The Keynesian theory of unemployment, shaped by the experience of the Great Depression, emphasizes the worker who lost his job when demand fell and who will only find a new job when aggregate demand increases sufficiently. In contrast, the best of the modern work on unemployment[2] has elaborated Stigler's (1961) analysis of search behavior with a model in which the unemployed worker samples job offers until he finds one that exceeds his optimal reservation wage. In this view, layoffs are characterized as involuntary while the return to work is regarded as the choice of the individual. A number of writers have suggested that the serious problem of unemployment is associated with

---

I am grateful to the National Science Foundation for financial support, to the University of California at Berkeley for the opportunity to return to this problem during my term as Ford Research Professor, and to Alan Auerbach and Robert Hall for comments on a previous draft.

[1] Specific evidence on the relative importance of temporary layoffs is presented in Sec. 1 below.

[2] This includes, e.g., Hall (1970, 1972), Holt (1970), McCall (1970), Mortensen (1970), Phelps (1970, 1972), Perry (1972), Tobin (1972), and Gordon (1973). This general view of unemployment can be traced back to Hicks (1932) and Pigou (1933).

[*Journal of Political Economy*, 1976, vol. 84, no. 5]

© 1976 by The University of Chicago. All rights reserved.

these apparently involuntary layoffs rather than with the substantial number of voluntary quits.

These theoretical formulations ignore the overwhelming importance of temporary layoffs. Most workers remain with a single firm for a very substantial period even though they may experience frequent spells of temporary unemployment. Most of those who are laid off know that they will soon return to their employer, protected by seniority arrangements and by their job-specific human capital [3] For them, the theory of job search is largely irrelevant [4] Moreover, in contrast to the search theory, it is the employers who determine the durations of these individuals' spells of unemployment

The traditional distinction between quits as voluntary and layoffs as involuntary also ceases to be valid Because workers remain with the same employer through several spells of unemployment, the frequency and duration of temporary layoffs must be regarded as part of the total package of compensation and conditions In a competitive labor market, employers will have to offer the economically feasible combination of unemployment, wages, and conditions *that workers prefer.* In collective-bargaining situations, the pattern of temporary layoffs may be an explicit part of the labor contract Although any particular layoff may be involuntary, the general pattern and rules are the result of explicit or implicit voluntary agreements Moreover, even particular layoffs may be voluntary Many collective-bargaining agreements contain "inverse seniority" provisions that give the most senior workers the opportunity to be laid off first and rehired last [5] Union contracts may also require that firms lay workers off instead of reducing average hours per week [6]

This paper will develop a theory of temporary layoffs Specific attention will be given to the question of why employment is reduced instead of hours The role of unemployment insurance and of taxes will be examined in detail

In an earlier study of unemployment,[7] I emphasized both temporary layoffs and unemployment insurance I explained there, and more fully in a subsequent paper (Feldstein 1974), that the impact of unemployment insurance is magnified by its interaction with the income tax. The general nature of this effect is clear, but a formal model is needed to indicate the order of magnitude of the impact and to provide a framework for future empirical research. The model of temporary layoffs that is developed in the current paper shows how unemployment insurance magnifies the effect on employment of changes in demand and increases the change in employment relative to the change in average hours. The analysis also shows how the unemployment insurance program can induce temporary layoffs or "unemployment holidays" even when there is no decrease in demand

---

[3] See Oi (1962), Becker (1964), and Parsons (1972) for analyses of the relation of specific human capital to the relatively permanent attachment of workers to individual firms

[4] There is some evidence that those who return to their original employer also investigate other jobs while they are out of work This may in part reflect a local unemployment insurance requirement that those who receive benefits must generally engage in active job search Whatever the reason for this search, the fact remains that most of those who are laid off do return and can expect to return to their original employer

[5] A study by the Bureau of Labor Statistics (U S Department of Labor 1972) describes the inverse seniority provisions in a large number of union contracts

[6] See, e g , the contract between the United Auto Workers and General Motors.

[7] Feldstein (1973*a*) Part of this study was reprinted in Feldstein (1973*b*).

The current paper focuses on the role of temporary layoffs as part of the higher "permanent" rate of unemployment[8] During normal times, random fluctuations in demand for specific products contribute to a continuously high rate of layoffs It is this phenomenon rather than occasional cyclical increases in unemployment that is the subject of the current analysis[9]

1. The Magnitude of Temporary Layoffs[10]

Although the purpose of this paper is to develop a theoretical analysis, it is useful to provide some empirical information about the relative importance of temporary layoffs

The National Longitudinal Survey that was conducted for the Department of Labor provides information on temporary layoffs among older men in all industries[11] In the survey year 1966–67, 9 percent of those surveyed experienced some unemployment while only 4 percent reported an involuntary job change[12] Since many of those who are unemployed have more than one spell per year and since some of those who are laid off will move directly into a new job without experiencing any unemployment, the ratio of job changes to total layoffs will be less than four-ninths Moreover, over the 3-year period 1966–69, only

---

[8] I have used the term "permanent" rate of unemployment rather than "natural" rate to emphasize that this equilibrium rate reflects taxes, unemployment insurance, and other institutional features of the labor market Any decrease in unemployment below this rate *through increases in aggregate demand* increases the rate of inflation In contrast, institutional changes can lower the unemployment rate without increasing inflation

[9] After the work for this paper was substantially complete, I had the opportunity to read Baily (1974*b*) He develops, especially in pp 16–23, an interesting alternative model of the impact of unemployment insurance There are too many differences in the structure of the models to permit direct comparison with the implications of the current model

[10] I have done extensive empirical analysis of temporary layoffs since this paper was written. Some basic estimates are presented in Feldstein (1976)

[11] The survey was directed by Dr Herbert Parnes and conducted by the Bureau of the Census The figures cited here were published in U S Department of Labor (1970) and Parnes, Nestel, and Andrisani (1972)

[12] U S Department of Labor (1970), pp 25 and 54

6 percent of those surveyed changed jobs involuntarily while there were at least 30 layoffs per 100 workers.[13]

A quite different type of information is provided by the question in the Current Population Survey about job search activities of the unemployed population To be counted as unemployed by this survey, an individual must have made specific efforts to find a job during the past 4 weeks unless he is waiting to be called back to a job from which he had been laid off or is waiting to report to a new job within 30 days. It is striking therefore that, in February 1975, more than 41 percent of the unemployed men aged 25–64 were not classed as job seekers. The unreported percentage of nonsearchers among those who had been laid off is of course much higher.[14]

There is further evidence that a very high proportion of workers laid off by manufacturing firms are subsequently recalled to the same employer During the past decade, there were approximately 1 33 layoffs per 100 employees each month in manufacturing firms [15] During the same period, accessions other than new hires by these firms averaged 1 11 per 100 employees per month, or 83 percent of the number of layoffs These accessions include only rehires and transfers between establishments within the same firms Although separate estimates of the numbers of rehires and transfers are not available, there is some evidence that the number of transfers is very small.[16] This implies that at least 75 percent of the workers laid off in manufacturing are subsequently rehired by their original employer

Although more detailed analysis of those who have been laid off would be desirable, the current data are sufficient to show the importance of temporary layoffs among workers who have long affiliations with their firms. I turn therefore to the theoretical analysis of this type of unemployment.

---

[13] See Parnes et al (1972), p 32, for the rate of involuntary job changes The layoff rate is a conservative estimate based on the number reporting some unemployment during each year

[14] U S Department of Labor (1975), p 30 These individuals were not "discouraged workers" who did not seek work because they believed that none was available; such individuals are not counted as part of the unemployed

[15] These figures on manufacturing turnover are published in *Employment and Earnings*, a monthly publication of the Department of Labor Layoffs are defined by the Department of Labor as "suspensions without pay lasting or expected to last more than 7 consecutive days, initiated by the employer without prejudice to the worker" (see, e g., U.S Department of Labor [1975], p 145) In addition to voluntary quits, other separations include death, retirement, permanent disability, entrance into the armed forces, transfers to other establishments of the same company, or discharges "for cause" (e.g , incompetence, dishonesty, etc) No information is available on the composition of these other separations.

[16] Telephone interviews with the individuals who prepared the turnover report for the Department of Labor at each of the six largest manufacturing employers in the Boston metropolitan area disclosed that reported transfers were always not greater than 5 percent of other accessions. Two firms did not even count transfers in their turnover figures.. Hamermesh (1969) also states that the flow of other accessions is composed mainly of rehires

2. A Model with Temporary Unemployment and Variable Hours

The current section presents a model of a firm that can adjust employment and hours in response to variations in the demand for its products These variations in demand are taken to be random fluctuations that are independent of the aggregate fluctuations in demand and in employment

The firm will be treated as a price taker in its product market. To emphasize the role of temporary layoffs, the firm's employees will be assumed to be "permanently" attached to the firm, that is, attached for the relevant planning period [17] For simplicity, a fixed capital stock that is appropriate for this number of employees will also be assumed If the labor market is competitive, the combination of wages, hours, and employment offered to these workers must produce as high a level of utility as they could obtain elsewhere Stated somewhat differently, a competitive labor market will result in that combination of wages, hours, and employment which maximizes the utility of the workers subject to the economic constraints imposed by demand, production, the cost of capital, and the relevant tax and transfer rules The current model will therefore pose the problem as one of maximizing the utility of the firm's "permanent" employees Treating the problem of optimal layoffs from the point of view of these employees themselves should underscore the inappropriateness of the traditional notion that such layoffs are involuntary

The analysis assumes all employees are identical and ignores employee risk aversion The absence of risk aversion seems a reasonable simplification, since the spells of unemployment are both temporary and brief.[18] With no risk aversion, the variation in demand can be regarded as periodic rather than stochastic [19] More specifically, I will study the behavior of employment and hours during a "year" when demand falls for a fraction λ of that year. Each worker bases his consumption during the year on his expected net income and uses his assets to smooth any difference between actual and expected income.[20]

---

[17] To avoid misunderstanding, I should emphasize that I regard this as an analytically useful limiting case and not as a description of reality A more realistic model of a firm would have to recognize a spectrum of workers of varying degrees of attachment

[18] The Current Population Survey estimated that the unemployed survey respondents in 1974 had been out of work for an average of 9 7 weeks This substantially overstates the average duration of completed spells of unemployment Kaitz (1970) estimated that the average duration of completed spells is only about 65 percent of the average duration to survey as estimated by the Current Population Survey The duration of temporary layoffs is also substantially shorter than the duration of permanent separations (see Feldstein 1976)

[19] That is, with no risk aversion, the individuals are concerned only with expected values of income and work The random period of reduced demand can be summarized by its expected duration

The current analysis assumes that employed workers are paid the same hourly wage rate when demand is low as when it is normal. This corresponds roughly to observed experience[21] and can be justified explicitly in the context of the current model Since employees in this model are affiliated with the firm "permanently" and since they use their personal assets to stabilize consumption when employment and hours vary, the employees (and the firm) are concerned only about the expected annual income and not about any pattern of within-year variation.[22]

This discussion is in sharp contrast to the analysis of Baily (1974*a*) and Gordon (1974), who show that wage rates do remain stable if workers are risk averse while firms are risk neutral The force of their analysis is substantially weakened if it is recognized that, because employees can use assets to smooth the level of consumption, the potential aversion to risk in the form of varying wages will be substantially reduced Moreover, the risk-avoidance logic of the Baily-Gordon model requires that firms stabilize *real* wages where the observed evidence is that money wages may adjust slowly to increasing or decreasing prices. The Baily-Gordon model is thus inadequate for the job of explaining the relatively stable money wages when prices are rising or falling. The actual reason for wage stickiness must be sought elsewhere in such things as the desire to avoid arbitrary wage changes and to choose the workers who will be permanently or temporarily laid off It is clear that the pattern of a nearly permanent attachment of each worker to a single firm must play a central role in the explanation of the observed wage stability

The firm's technology will be represented by a production function that recognizes the important distinction between increasing the number of employees and increasing the average hours per employee

$$X = G(K, N, h), \tag{2 1}$$

---

[20] A worker who has 8 weeks of unemployment in a year will find that, because of income taxes and unemployment compensation, his annual earnings will fall by substantially less than 10 percent In families with two earners, total family income will fall proportionately less. Surveys of the unemployed indicate that dissaving and postponing payments are used to maintain the level of consumption (see, e g., Cohen [1960] and Lester [1962]) This weak "permanent-income" assumption, i e , that workers maintain an effectively constant level of consumption, is in sharp contrast to recent models of Baily (1974*a*) and Gordon (1974) that emphasize risk aversion and that assume either that employees do not have sufficient assets to smooth their consumption or that they base consumption on current income rather than permanent income For the analysis of temporary unemployment, the current assumption seems preferable.

[21] In practice, the average hourly earnings of a worker may fall when demand is low because overtime is eliminated and because workers are "bumped down" into lower-grade jobs

[22] Recall that employees are not assumed to be risk averse or, more precisely, that the effects of risk aversion are small enough to be ignored

where X is the current annual rate of output, N is the number of employees, h is the average hours per employee, and K is the capital stock The traditional specification of the labor input as man-hours $(h \cdot N)$ ignores the fact that an increase in average hours per man also increases the flow of capital services This suggests the specification $X = G(hK, hN)$ or, with constant returns to scale, $X = hG(K, N)$. In practice, a variation in average hours will have a more complex effect Because of the fixed time involved in starting and stopping work each day, the number of effective working hours will increase more than proportionately to the number of hours for which a worker is paid Offsetting this is the greater fatigue that accompanies long hours It is best, therefore, to use the unrestricted form of equation (2 1) [23]

The variation in the demand for the firm's product can be described by a fall in the price that the firm receives When demand is normal, the price is p_0 During a fraction λ of the year, the price falls to p_1 [24] The total annual revenue of the firm is thus

$$(1 - \lambda)p_0X_0 + \lambda p_1X_1 = (1 - \lambda)p_0 \; G(K, N_0, h_0) + \lambda p_1 \; G(K, N_1, h_1). \tag{2 2}$$

Note that both the original number of employees (N_0) and the capital stock (K) is fixed

Workers who are laid off when demand falls receive unemployment insurance benefits at an annual rate of b Total unemployment benefits per year are therefore

$$B = \lambda b(N_0 - N_1) \tag{2 3}$$

In the model, as in the United States, these benefits are financed by a tax on firms that is related to the previous benefits collected by the firm's employees This system of "experience rating" is very imperfect; many firms pay a flat rate of tax per employee that is unaffected by marginal changes in unemployment [25] This imperfect experience rating will be represented in the current steady-state context by making the tax a fraction e of annual benefits plus a constant T. Thus the unemployment tax is $eB + T$.[26]

---

[23] In Feldstein (1967) I estimated a Cobb-Douglas production function with separate coefficients of men and hours using an aggregate cross section of industries. The results implied that the elasticity of output with respect to average hours exceeds the sum of the elasticities with respect to the number of employees and the capital stock

[24] This model of price variation can also be regarded as an approximation to a more realistic description in which firms respond to lower demand by increasing sales effort or accumulating inventories These cost increases are, for the current purpose, equivalent to a reduction in price The model can be modified to allow for a price-setting firm that experiences a shifting demand curve, price setting does not affect any conclusions of the analysis

[25] Although the exact provisions of the law differ among the states, the usual pattern is a tax rate that varies among firms on the basis of each firm's past experience More specifically, the tax rate is a decreasing function of the difference between its cumulative tax payments (since the beginning of the unemployment insurance program) and the cumulative benefits paid to its employees (including former employees) A crucial feature of the tax schedule is that there is a relatively low maximum tax rate and a positive minimum tax rate As a result, many firms with high unemployment rates have "negative balances" in their accounts, i e , have paid less in taxes than their employees

Since markets are assumed here to be competitive, the firm's total revenue must be exhaustively divided among the cost of capital services[27] (cK), the aggregate wage bill (W), and the unemployment tax ($eB + T$). From (2.2), this implies

$$cK + W + eB + T = (1 - \lambda)p_0G(K, N_0, h_0) + \lambda p_1G(K, N_1, h_1). \tag{2.4}$$

The total income of the N_0 permanent employees[28] is the sum of the aggregate wage bill and the unemployment benefits, $W + B$. If wage income is taxed at rate t_y while unemployment benefits are taxed at rate t_b,[29] total net income is

$$Y = (1 - t_y)W + (1 - t_b)B \tag{2.5}$$

Before the year begins, each employee has an equal probability of becoming unemployed. For each employee, the expected net income is thus

$$y = (1 - t_y)\frac{W}{N_0} + (1 - t_b)\frac{B}{N_0} \tag{2.6}$$

Since there is no risk aversion, this expected income is a sufficient description of the employee's stochastic income prospect.

The utility of the representative employee will be represented by a strongly separable function. Utility at each point in time depends only on the level of consumption and the amount of work done at that time. In addition, the instantaneous utility function is separable so that the number of working hours does not affect the utility of consumption and

---

have received in benefits. These firms with high unemployment rates face the maximum tax rate, an increase in layoffs causes no increase in tax payments. Similarly, the large number of firms with substantial "positive balances" face the minimum rate and would continue to do so even if their rate of layoffs increased, as long as that increase was not too great. Becker (1972) presents extensive evidence of the importance of the firms that pay no effective marginal tax. For example, in New York in 1967 some 59 percent of all benefits were related to firms with negative balances while 28 percent of firms paid the minimum tax. Similarly, in Massachusetts, 57 percent of benefits were related to firms with negative balances while 18 percent of firms paid the minimum tax.

[26] For firms at the maximum or minimum tax rates, $e = 0$. The evidence in the previous footnote that more than half of the benefits are paid to such firms in Massachusetts and New York implies that the average of the e values is less than one-half in those states.

[27] This includes both the interest cost and depreciation. It is defined as the net after-tax cost of capital services.

[28] Exclusive of any income from assets that they own. Total income refers to the total income derived from employment by the firm.

[29] In the United States, $t_b = 0$ if sales taxes are ignored. In Canada and some other countries, t_b is the rate of personal income tax (t_y minus the payroll tax rate).

the level of consumption does not affect the utility of leisure. Since consumption is constant during the year and independent of whether the individual becomes unemployed, the utility function can be written[30]

$$U = u(y) + (1 - \lambda)\ v(h_0) + \lambda \frac{N_1}{N_0} v(h_1) + \lambda \frac{N_0 - N_1}{N_0} v(0) \qquad (2\ 7)$$

Because of the competitive labor market, the firm's problem is to select the values of h_0, h_1, and N_1 that maximize this utility of the representative employee, subject to the constraint implied by equations (2 1)–(2 6) The properties of this optimum will now be examined

3. The Effects of Changes in Demand

The purpose of this section is to assess how a change in the demand price (p_1) affects employment (N_1) and average hours (h_1) The sensitivity of this response to the unemployment insurance and tax parameters will be examined. The analysis shows that the current unemployment insurance greatly magnifies the effect on employment of changes in demand and reduces its effect on hours The analysis also shows that the unemployment insurance program may induce unemployment even if there is no fall in demand

The optimum policy of employment and hours is characterized by the three first-order conditions for the maximization of utility in equation (2 7) [31]

$$u'(y)\ \frac{\partial y}{\partial N_1} + \frac{\lambda}{N_0}\,[v(h_1) - v(0)] = 0, \qquad (3\ 1)$$

$$u'(y)\ \frac{\partial y}{\partial h_1} + \lambda\left(\frac{N_1}{N_0}\right) v'(h_1) = 0, \qquad (3\ 2)$$

and

$$u'(y)\ \frac{\partial y}{\partial h_0} + (1 - \lambda) v'(h_0) = 0 \qquad (3\ 3)$$

---

[30] The utility of consumption is written $u(y)$ since consumption depends on the steady-state net income y The separability of the instantaneous utility function implies that

$$\begin{aligned} u(y) &= (1 - \lambda)\ u(y \mid h = h_0) + \lambda(N_1/N_0)\ u(y \mid h = h_1) \\ &\quad + \lambda[(N_0 - N_1)/N_0]\ u(y \mid h = 0) \\ &= \{(1 - \lambda) + \lambda(N_1/N_0) + \lambda[(N_0 - N_1)/N_0]\}u(y) \end{aligned}$$

Recall that I am examining the polar case in which layoffs are temporary, so that there is no permanent separation and no job search

[31] There are also two further constraints that $N_1 \leq N_0$ and that both h_1 and h_0 not exceed some upper limit These constraints will generally not be binding and can therefore be ignored Exceptions will be noted below

From equations (2.2)–(2.6), the derivatives of average net income with respect to employment and hours are

$$\frac{\partial y}{\partial N_1} = \frac{\lambda(1 - t_y)}{N_0}[p_1 G_{N1} + eb] - \frac{(1 - t_b)b\lambda}{N_0}, \tag{3.4}$$

$$\frac{\partial y}{\partial h_1} = \frac{\lambda(1 - t_y)}{N_0} p_1 G_{h1}, \tag{3.5}$$

and

$$\frac{\partial y}{\partial h_0} = \frac{(1 - \lambda)(1 - t_y)}{N_0} p_0 G_{h0}, \tag{3.6}$$

where $G_{N1} = \partial G/\partial N_1$, etc. The derivatives of income with respect to h_1 and h_0 are simply the marginal revenue product per employee after income tax. In contrast, the derivative of income with respect to employment (N_1) must be adjusted for the difference between the forgone unemployment insurance benefits $[(1 - t_b)b\lambda/N_0]$ and the incremental net cost of additional experience-rated tax that would have to be paid $[\lambda(1 - t_y)eb/N_0]$.

a) Optimal Employment with Fixed Hours

Before examining the implications of the conditions for optimal average hours, it is useful to consider the behavior of employment if hours are arbitrarily fixed. Substituting equation (3.4) into equation (3.1) and rearranging terms yields the first-order condition

$$G_{N1} = \left[\frac{[(1 - t_b) - (1 - t_y)e]b}{(1 - t_y)} + \frac{v(0) - v(h_1)}{(1 - t_y)u'(y)}\right]p_1^{-1} \tag{3.7}$$

This implies that the optimal marginal product of an additional employee during the period of reduced demand varies inversely with the price of the product. Since this marginal product (G_{N1}) varies inversely with employment (N_1), equation (3.7) implies that a fall in p_1 decreases N_1.[32] The magnitude of the change in N_1 depends on two factors. (1) The net unemployment insurance subsidy,

$$J = \frac{[(1 - t_b) - (1 - t_y)e]b}{(1 - t_y)}, \tag{3.8}$$

is the value, measured in terms of equivalent pretax earnings, of the excess of the net unemployment benefits $[(1 - t_b)b]$ over the net cost of the experience-rated tax $[(1 - t_y)eb]$ (2) The value of leisure,

---

[32] It is possible to satisfy (3.7) when p_1 falls without a reduction in N_1 if h_1 increases sufficiently. I return to this below. Note that (3.7) is a generalization of the textbook case in which $b = t_y = v(0) = 0$ implies $p_1 G_{N1} = -v(h_1)/u'(y)$, that is, the marginal revenue product of labor equals the disutility of hours divided by the marginal propensity of money.

$$L(h_1) = \frac{v(0) - v(h_1)}{(1 - t_y)u'(y)}, \tag{3 9}$$

is the value, measured in terms of pretax earnings, of the utility difference between no-work $[v(0)]$ and working h_1 hours $[v(h_1)]$

It is useful to consider the empirical magnitudes of these two terms in order to have a sense of their potential significance in the analysis that follows Consider first the unemployment compensation term As I noted above, $t_b = 0$ in the United States In the large fraction of cases in which there is no effective experience rating, $e = 0$ and the unemployment insurance subsidy is $J = b/(1 - t_y)$, that is, the value of unemployment benefits in terms of equivalent taxable wages In a detailed study of unemployment insurance in each state (Feldstein 1974), I found that the average value of $b/(1 - t_y)$ is approximately two-thirds of the average gross wage For those firms in which there is full experience rating, $e = 1$ and the unemployment insurance subsidy is $J = t_y b/(1 - t_y)$. The relevant marginal income tax rate (t_y) includes not only the federal and state income tax rates but also the employer and employee contributions for social security, a conservative estimate of the relevant marginal rate for unemployment insurance recipients is one-third[33] This implies a value of $t_y b/(1 - t_y)$ equal to two-ninths of the gross wage If half of the unemployment benefits are paid to those with $e = 0$ and half to those with $e = 1$, the overall average of the unemployment insurance subsidy is approximately four-ninths, or nearly 50 percent of the corresponding gross wage

The relative value of leisure, $[v(0) - v(h_1)]/(1 - t_y)u'(y)$, is harder to assess For many, such leisure has little or no value For others, a short spell of leisure may be worth as much as the lost earnings Gordon (1973) estimated that the value of waiting time and search time during unemployment was equal to only 53 cents per hour for 1971 Because search is irrelevant in our analysis, this represents an overestimate of the value of waiting time during temporary layoff But even 53 cents per hour for 40 hours per week and a marginal income tax rate of $t_y = 1/3$ imply a value of $[v(0) - u(h_1)]/(1 - t_y)u'(y) = \$32$, or less than 40 percent of the corresponding gross wage[34] This calculation suggests that, on average, the unemployment compensation subsidy exceeds the value of leisure and, in firms where there is no marginal experience rating, the subsidy may be nearly twice the value of leisure. This implies that, on average, the unemployment insurance subsidy more than doubles the impact of a change in p_1 on the optimal marginal product of employment. Where there is no marginal experience-rated tax ($e = 0$), unemployment insurance increases the change in G_N to more than three times what it would otherwise be.

---

[33] For temporary layoffs, the relevant tax includes the employer's social security tax as well as the employee's Any family that pays federal income tax therefore has a combined marginal rate of at least 26 percent In Massachusetts, a family with \$4,000 of taxable income would pay a combined rate (including state income tax) of 33 percent

[34] The average 1971 gross weekly wage in private nonagricultural employment was \$127 28 Stein (1963) estimated that the unemployed earn substantially less than all employees, Gordon (1973) updated his figures and concluded that the unemployed had previously earned only 69 percent of the average wage This implies a wage of \$88 and therefore a relative value of leisure of 39 percent

With these order-of-magnitude calculations as background, let us consider the implications of (3.7) in more detail. Note first the obvious implication that it is optimal to maintain full employment ($N_1 = N_0$) if the unemployment program provides no net subsidy [$(1 - t_b)b = (1 - t_y)eb$] and there is no disutility of work [$v(0) = v(h_1)$]. With these conditions, the right-hand side of equation (3.7) is zero. Taken literally, this implies that employment should be expanded indefinitely until the marginal product of labor is driven to zero. Correctly interpreted, this implies expansion of unemployment until the constraint $N_1 \leq N_0$ becomes binding.

The opposite extreme, in which it is desirable to lay off all the employees during the period of reduced demand, is also interesting to examine. Let $G_{N1}(0)$ be the marginal product of the last employee.[35] Then $N_1 = 0$ is optimal if

$$p_1 G_{N1}(0) < \frac{[1 - t_b - (1 - t_y)e]b}{(1 - t_y)} + \frac{v(0) - v(h_1)}{(1 - t_y)u'(y)}, \qquad (3.10)$$

that is, if the marginal revenue product is less than the combined value of net unemployment insurance benefits and the income equivalent of the leisure. As I noted above, when $e = 0$ the unemployment insurance subsidy is equal to approximately two-thirds of the regular gross wage. Since this greatly increases the value of the right-hand side, it makes it much more likely that it would be optimal to close the plant temporarily.

Totally differentiating equation (3.7) with respect to N_1 and p_1 for given J yields (with $G_{NN} = \partial^2 G/\partial N_1^2$)

$$\frac{dN_1}{dp_1} = \frac{-G_N - [v(0) - v(h_1)]/[(1 - t_y)u'(y)] \quad [u''(y)/u'(y)] \cdot (\partial y/\partial p_1)}{p_1 G_{NN} + [v(0) - v(h_1)]/[(1 - t_y)u'(y)] \quad [u''(y)/u'(y)] \cdot (\partial y/\partial N_1)} \qquad (3.11)$$

If $u'' = 0$, it follows directly from the technological properties ($G_N > 0$ and $G_{NN} < 0$) that $dN_1/dp_1 > 0$, that is, that a fall in demand will decrease employment. If $u'' < 0$, the second term is also negative so that the conclusion is unchanged [36] The analysis that follows will be simplified by assuming a constant marginal utility of income, $u' > 0$ and $u'' = 0$.

---

[35] Or, to be more realistic, the average product of the smallest number of employees who can technologically be employed.

With this assumption, it also follows from totally differentiating (3 7) with respect to N_1 and J that

$$\frac{dN_1}{dJ} = \frac{1}{p_1 G_{NN}} < 0 \tag{3 12}$$

Increasing the unemployment insurance subsidy reduces employment at any level of reduced demand

An example based on a Cobb-Douglas technology will illustrate the order of magnitude of this effect Since h_1 as well as K is temporarily regarded as fixed, the Cobb-Douglas production function can be written $G = aN_1^{\alpha}$ The first-order condition of equation (3 7) thus becomes

$$\alpha a N_1^{\alpha-1} = \left[\frac{[(1 - t_b) - (1 - t_y)e]b}{(1 - t_y)} + \frac{v(0) - v(h_1)}{(1 - t_y)u'(y)}\right] p_1^{-1} \tag{3 13}$$

or

$$N_1 = (\alpha a)^{1/(1-\alpha)} \left[\frac{[(1 - t_y) - (1 - t_y)e]b}{(1 - t_y)} + \frac{v(0) - v(h_1)}{(1 - t_y)u'(y)}\right]^{1/(1-\alpha)} p_1^{1/(1-\alpha)} \tag{3 14}$$

The Cobb-Douglas technology therefore implies a constant elasticity of employment with respect to the price during the period of reduced demand

Of greater interest is the effect of the unemployment insurance subsidy on this level of unemployment Note that at any value of p_1 there is a constant elasticity relation between N_1 and the sum of the insurance subsidy and the value of leisure

$$\frac{J + L(h)}{N_1} \; \frac{\partial N_1}{\partial(J + L(h))} = \frac{-1}{1 - \alpha} \tag{3 15}$$

This implies that the current unemployment insurance subsidy is likely to have a substantial effect on the level of employment during the period of reduced demand Most empirical estimates place α at two-thirds or higher This makes the elasticity of N_1 with respect to $J + L(h)$ equal to -3 or greater I noted above that J is probably greater than $L(h)$ on average and may be twice as great for firms in which there is no effective experience rating ($e = 0$) Even $J = L(h)$ implies that the current unemployment insurance subsidy doubles $J + L(h)$ and, with an elasticity of -3, reduces N_1 to one-eighth of the value that would be optimal with no unemployment insurance subsidy. The effect of the subsidy is thus to reduce optimal employment drastically.[37]

---

[36] This assumes that $\partial y/\partial N_1 > 0$, i e , that more employment increases average income This must be true at an optimum, since more employment reduces leisure. Although for particular individuals $\partial y/\partial N_1 < 0$ because the unemployment insurance subsidy exceeds the net wage (see examples in Feldstein [1974]), for the firm's employees as a whole this will not be true

It is interesting also to examine the elasticity of N_1 with respect to J for $J > 0$. Equation (3.15) implies

$$\frac{J}{N_1}\frac{\partial N_1}{\partial J} = -\frac{1}{1-\alpha}\,\frac{J}{J+L(h_1)}. \qquad (3.16)$$

With $\alpha = 2/3$ and $J = L(h)$, this yields an elasticity of employment during the period of reduced demand with respect to the unemployment insurance subsidy of -1.5. From the definition in equation (3.8), it is clear that a 10 percent increase in the benefit level implies a 10 percent increase in J and therefore a 15 percent decrease in employment. Making unemployment compensation taxable like ordinary income would reduce J by at least $100t_y$ percent,[38] with $t_y = 0.3$, J is cut by at least 30 percent and therefore N_1 increases by at least 45 percent.[39] Similarly, introducing experience rating (raising the value of e from zero to $e > 0$) reduces J by $100(1 - t_y)e/(1 - t_b)$, with $t_b = 0$, $t_y = 0.3$, and $e = 1$, J falls by 70 percent and optimal employment doubles.[40]

It should be emphasized that these conclusions rest on the assumption that working hours are institutionally fixed, unable to respond to changes in demand, and unaffected by the existence of the unemployment insurance subsidy. It is now time to examine the implications of dropping that assumption.

b) Optimal Working Hours and Employment

When hours as well as the number of workers can be varied, the response to a fall in demand is much more complex. It may be optimal for the firm to reduce either employment or hours or both, and, more surprisingly, it may be optimal for the firm to increase either one of these.[41] The optimal behavior will depend on the properties of the production function and on the individuals' preferences for leisure This section will examine this ambiguity and will assess the effect of the unemployment insurance subsidy on both employment and hours

---

[37] Note that equation (3.11) shows that $N_1 < N_0$ even if $J = 0$. Two caveats should also be noted. First, the average condition that J is approximately equal to $L(h)$ may not be representative of any particular firm. For those firms with $e = 1$, I noted above that J is approximately two-ninths of gross wages or approximately half of the value of $L(h)$ assumed here. For such firms, the existence of $J > 0$ reduces N_1 by a factor of $(9/11)^3 = 0.55$. In contrast, for firms with $e = 0$, the effect is even greater than the reduction indicated in the text. Second, the implied changes in employment do not refer to all of the employees of a firm but only to those who produce that particular product for which demand has temporarily fallen. Temporary layoffs for particular types of workers are consistent with an observed relatively high overall level of employment.

[38] With $e = 0$, J is reduced from $b/(1 - t_y)$ to b, thus implies a proportionate reduction of t_y. With $e > 0$, the reduction in J is larger.

[39] This increase in N_1 may make $N_1 > N_0$ and therefore imply full employment.

[40] See n. 39 above.

[41] Of course, the constraint $N_1 \leq N_0$ may be binding and may thus induce a decrease in h_1 that would not otherwise be optimal. Similarly, if there is a constraint that permits hours to be reduced but not increased ($h_1 \leq h_0$), this constraint may be binding, the problem is then equivalent to the fixed-hours case of Sec. 3.1. For a different model of the determination of working hours with special reference to overtime, see Flemming (1974).

The first-order condition for average hours during the period of reduced demand is, from equations (3 2) and (3 5),

$$G_{h1} = -\frac{N_1 v'(h_1)}{(1 - t_y)u'(y)p_1} \qquad (3\ 17)$$

Since G_{h1}/N_1 is the marginal product per employed worker of increasing average hours, equation (3 17) states that the utility of the after-tax marginal revenue product per employed worker $[u'(y)(1 - t_y)p_1 G_{h1}/N_1]$ is equal to the marginal utility of additional leisure $[-v'(h_1)]$ Note that, although the equilibrium condition is not directly affected by the unemployment insurance subsidy, it is affected indirectly Since the subsidy changes N_1, it directly alters the right-hand side of (3 17) and also changes the values of G_{h1} at every level of h_1 The net effect of the subsidy on optimal hours can only be determined by examining the joint optimum of hours and employment

First, however, it is necessary to consider the optimum average hours during the period of normal demand From equations (3 3) and (3 6), the first-order condition is

$$G_{h0} = -\frac{N_0 v'(h_0)}{(1 - t_y)u'(y)p_0} \qquad (3\ 18)$$

The interpretation of this equation is also straightforward the utility of the after-tax marginal revenue product of increasing average hours must equal the marginal utility of additional leisure The most important implication is that, if the marginal utility of income is constant $[u''(y) = 0]$, the optimal number of hours during the period of normal demand is not affected by the magnitude of the unemployment insurance subsidy or by the response to the fall in demand This decomposition of the optimum conditions permits us to focus on the two first-order conditions for N_1 and h_1, equations (3 7) and 3 17)

To assess the effects of changes in demand (p_1) and in the unemployment insurance subsidy (J), we totally differentiate these equations and obtain

$$\begin{aligned}(1 - t_y)u'(y)p_1 G_{NN}\, dN_1 &+ [(1 - t_y)u'(y)p_1 G_{Nh} + v'(h_1)]\, dh_1 \\ &= (dJ - G_N\, dp_1)(1 - t_y)u'(y) \qquad (3\ 19)\end{aligned}$$

and

$$\begin{aligned}[(1 - t_y)u'(y)p_1 G_{Nh} &+ v'(h_1)]\, dN_1 \\ &+ [(1 - t_y)u'(y)p_1 G_{hh} + N_1 v''(h_1)]\, dh_1 \\ &= -(1 - t_y)u'(y)G_h\, dp_1 \qquad (3\ 20)\end{aligned}$$

Consider first the effect of a change in the unemployment insurance subsidy at any level of reduced demand (p_1). Solving these two equations yields

$$\frac{dN_1}{dJ} = \frac{p_1 G_{hh} + N_1 v''(h_1)(1 - t_y)^{-1}[u'(y)]^{-1}}{\Delta} \tag{3.21}$$

and

$$\frac{dh_1}{dJ} = \frac{-p_1 G_{Nh} - v'(h_1)(1 - t_y)^{-1}[u'(y)]^{-1}}{\Delta}, \tag{3.22}$$

where Δ is the determinant of the coefficients of equations (3.19) and (3.20) and is positive as a necessary second-order condition for the maximization of utility in equation (2.7).

Since $(1 - t_y)u'(y)p_1 G_{hh} + N_1 v''(h_1) < 0$ is also a necessary second-order condition, dN_1/dJ is unambiguously negative. The result of the previous section, that an increase in the unemployment insurance subsidy decreases employment, remains true even when hours are allowed to adjust. Moreover, since $(1 - t_y)u'(y)p_1 G_{Nh} + v'(h_1) < 0$,[42] equation (3.22) shows that the unemployment insurance subsidy increases the average hours per worker during the period of reduced demand. Thus, unemployment insurance exacerbates the effect of demand on employment and increases the incentive to use layoffs rather than shorter hours for any given reduction in man-hours.

Although the effect of the unemployment insurance subsidy is unambiguous, the effect of a change in demand cannot be determined without further restriction. Equations (3.19) and (3.20) imply

$$\begin{aligned}\frac{dN_1}{dp_1} = {} & \frac{-[(1 - t_y)u'(y)p_1 G_{hh} + N_1 v''(h_1)]G_{N1}}{\Delta} \\ & + \frac{[(1 - t_y)u'(y)p_1 G_{Nh} + v'(h_1)]G_{h1}}{\Delta}\end{aligned} \tag{3.23}$$

and

$$\begin{aligned}\frac{dh_1}{dp_1} = {} & \frac{[(1 - t_y)u'(y)p_1 G_{Nh} + v'(h_1)]G_{N1}}{\Delta} \\ & - \frac{(1 - t_y)u'(y)p_1 G_{NN} G_{h1}}{\Delta}\end{aligned} \tag{3.24}$$

Since the denominators are positive, the signs of the optimal changes in N_1 and h_1 depend on the relative magnitudes of the positive and negative

---

[42] It is easy to show that this inequality is true. Recall from equation (3.17) that $(1 - t_y)u'(y)p_1 G_h + N_1 v'(h_1) = 0$. Dividing by $N_1 > 0$ and recognizing that $G_h/N > G_{hN}$ follows from the decreasing returns to variations in employment alone yield

$$(1 - t_y)u'(y)p_1 G_{hN} + v'(h_1) < (1 - t_y)u'(y)p_1 G_h N_1^{-1} + v'(h_1) = 0$$

as required.

terms in the numerator. This depends on both the technological conditions and the employees' preferences for leisure. Without further restriction, it is possible to eliminate only the combination $dN_1/dp_1 < 0$ and $dh_1/dp_1 < 0$, that is, that a decrease in price raises both employment and hours.[43] Of course, $dN_1/dp_1 < 0$ violates the restriction $N_1 \leq N_0$ and actually results in a change in h_1 only. *What can be said unequivocally is that the unemployment insurance subsidy increases the likelihood of a fall in employment and magnifies the size of any unemployment that occurs.*

Finally, it is interesting to note that, even with no fall in demand, the presence of the unemployment subsidy makes it optimal to have a period of feigned unemployment, an unemployment insurance holiday. This is easily seen in the basic first-order condition for employment, equation (3.7). Even if $p_1 = p_0$, the possibility of obtaining a subsidy at rate J for some part of the year ($J > 0$) implies that it is optimal to increase G_N during those weeks, that is, to reduce employment. Moreover, equation (3.21) shows that $dN_1/dJ < 0$ even if $p_1 = p_0$, that is, any increase in the unemployment insurance subsidy increases the optimal unemployment insurance holiday.

4. Some Further Implications

The analysis of Sections 2 and 3 is essentially microeconomic, focusing on the behavior of the individual firm and its employees. The macroeconomic implications of this pattern of temporary layoffs and the scope for a reform of unemployment insurance to achieve macroeconomic goals will be considered briefly in this section.

a) *Macroeconomic Effects of Unemployment Insurance*

The analysis in this paper implies that, under the economic conditions that have prevailed in the postwar period, our current system of unemployment compensation is likely to have increased the average rate of unemployment. The usual presumption, that unemployment compensation reduces unemployment because it automatically increases government spending when unemployment rises, is really irrelevant. When the economy is experiencing the average high "permanent" rate of unemployment, that is, when there is neither a cyclical peak nor a cyclical trough, the unemployment insurance system collects as much in tax revenue as it pays out in benefits. Although benefits do exceed tax revenues during cyclical downturns, the same fiscal stimulus would be provided through other expenditure increases or tax cuts by a government committed to maintaining aggregate demand The primary effect of our current system of unemployment compensation is not its contribution to aggregate demand but its adverse impact on the incentives of employers and employees

---

[43] Note that $dN_1/dp_1 < 0$ implies $-G_N[(1 - t_y)u'(y)p_1G_{hh} + N_1v''(h_1)] < -G_H[(1 - t_y)u'(y)p_1G_{Nh} + v'(h)]$ while $dh_1/dp_1 < 0$ implies $-G_H(1 - t_y)u'(y)p_1G_{NN} < -G_N[(1 - t_y)u'(y)p_1G_{Nh} + v'(h)]$. Since all four terms are positive, this implies that the product of the two smaller parts of the inequalities is less than the product of the two larger parts. After factoring out G_NG_H, this is equivalent to $\Delta < 0$, which is false. Therefore, dN_1/dp_1 and dh_1/dp_1 cannot both be negative.

Section 3 shows explicitly how the unemployment insurance subsidy increases the change in unemployment that follows any change in demand [44] If the subsidy were reduced to zero, the induced temporary unemployment would be eliminated. At this point, the reader may well be puzzled if demand falls by as much but there is a smaller reduction in employment, will there not be extra output that is not purchased?[45] The puzzle is easily solved Note first that it is clear that an increase in the supply of labor will, by Walras's law, result in demand for the additional output unless monetary policy does not provide adequate support for the new level of income What is the microeconomic process by which this occurs?

Clearly, if the firm's demand fluctuates during the year and its rate of output is nevertheless determined by a higher level of employment than is optimal for the periods of reduced demand, the firm will have to maintain an inventory that increases during the periods of reduced demand In order to reduce the inventory when demand returns to the "normal" level,[46] the full employment output of the firm must be less than the normal level of demand The immediate consequence of this is to require the firm to reduce the permanent size of its labor force below the current level of N_0 Two things should be noted First, the workers who are dismissed in this process would find jobs in other firms, this is a consequence of the full employment assured by Walras's law Second, and more important, if every firm shares proportionately in the increased demand that results from increased employment, each firm will be able to retain its original N_0 workers and to employ them with a lower average rate of unemployment Any actual dismissals (and hires by other firms) would occur only because of the second-order effects of a nonproportional expansion of demand In short, there is no problem in achieving the additional demand required to support the increased supply of labor

The current analysis is clearly related to recent evidence that "the Phillips curve has shifted outward" or, alternatively, that the "critical level of unemployment at which the rate of inflation begins to accelerate has increased."[47] There has been much emphasis on the changing demographic structure of the labor force as a reason for this shift of the Phillips curve. However, these demographic changes do not account fully for this shift. Moreover, these calculations fail to reflect other changes in labor force composition that would lead to lower unemployment rates, for example, the growing share of college-educated employees or the increasing share of workers who are employed in the government sector.

---

[44] Let me emphasize again that it is the unemployment insurance *subsidy* ($J > 0$) and not unemployment insurance itself ($b > 0$) that causes the increase in temporary layoffs

[45] It is of course true that a fall in average hours caused by this reduced unemployment will reduce output But the problem remains if hours are institutionally fixed The solution to be described is valid in either case.

[46] I assume here, as in the previous sections, that demand varies between "normal" and "low "

The increasing unemployment insurance subsidy is a potentially important cause of the rising trend in unemployment rates and the apparent shift of the Phillips curve Maximum weekly benefits have been rising more rapidly than wages during the past decade More significantly, the relevant marginal tax rate (t_y) paid by those who are likely to become unemployed has also risen rapidly, especially because of the growth of the social security tax (from a rate of 6 0 percent on the first $4,800 of earnings in 1960 to a rate of 11 7 percent on the first $14,100 in 1975) and of state and local income taxes The impact of this growing subsidy on the Phillips curve clearly deserves further empirical analysis

5. Conclusion

The overwhelming importance of temporary layoffs requires a revision of the basic conceptual framework within which layoffs and unemployment are now discussed This paper shows that the usual dichotomy between quits as voluntary separations and layoffs as involuntary is often wrong The common model of job search is relevant to understanding the behavior of those who quit their last job or who are entering the labor force, but it is not generally helpful for analyzing the experience of those who are laid off For most of this group, unlike the searchers, the return to work is not the result of a voluntary decision by the employee but of recall by the employer

And yet, as I stress by the technique of using a model that maximizes the utility of the representative employee, the optimal pattern or rules for temporary layoffs are part of the voluntary package of wages, hours, and work-sharing rules that employees choose or for which they bargain Any particular spell of unemployment may be both involuntary and loudly protested, even though the decision rule that led to the layoff may have been chosen by the employees [48]

---

[47] I put these two formulations in quotation marks to avoid taking sides in the debate over whether there is a nonvertical Phillips curve over at least some of the range. I will use the two formulations interchangeably in what follows. Phelps (1972) summarizes the evidence that the curve has been shifting, while Hall (1974) provides an estimate that that critical level of unemployment for a stable level of inflation has reached 5.5 percent.

[48] In using this broad definition of voluntary unemployment, I do not wish to imply that there is no involuntary unemployment. There are clearly circumstances in which, by any definition, there is involuntary unemployment

The employees' preference for the existing pattern of temporary layoffs is, of course, a preference that has been shaped by the unemployment insurance subsidy. There is nothing natural or optimal about the resulting high rate of unemployment. The current analysis shows that the unemployment insurance subsidy has a potentially very large impact on the rate of unemployment, causing layoffs when they would not otherwise happen and substantially magnifying the size of the layoffs that do occur It is important to emphasize that it is not unemployment insurance per se but the subsidy that it currently entails that causes this adverse distortion. If there were full experience rating in the unemployment insurance tax ($e = 1$) and if benefits were taxed like other income ($t_b = t_y$), the subsidy to temporary layoffs would be eliminated ($J = 0$) Unemployment insurance could then continue to provide protection for those who are temporarily laid off without any distortion in their behavior [49]

About half the adult unemployed at any time have been laid off. Among men aged 25–64, the fraction of layoffs rises to nearly three-fourths. The evidence presented in Section 1 of this paper indicates that most of this group of unemployed return to their original employers after a short spell of unemployment The model of temporary layoffs derived here indicates that the unemployment insurance subsidy may be responsible for a very high fraction of such layoffs If these conclusions are supported in future research, they will point the way to a reform of unemployment insurance that could substantially lower the permanent rate of unemployment without reducing the protection that is available to those who are without work

References

Baily, M N "Wages and Employment under Uncertain Demand" *Rev Econ Studies* 41, no 125 (January 1974) 37–50 (*a*)

———. "Unemployment and Unemployment Insurance" Yale Discussion Paper, 1974 (*b*)

Becker, Gary *Human Capital A Theoretical and Empirical Analysis with Special Reference to Education* New York Columbia Univ Press, 1964

Becker, Joseph *Experience Rating in Unemployment Insurance* Baltimore· Johns Hopkins Press, 1972.

Cohen, W J , Haber, W., and Mueller, E. *The Impact of Unemployment in the 1958 Recession* Special Committee on Unemployment Problems, U.S. Senate, Committee Print, June 1960

Feldstein, M S "Specification of the Labor Input in the Aggregate Production Function." *Rev. Econ Studies* 34 (October 1967) 375–86

———. *Lowering the Permanent Rate of Unemployment.* A study prepared for the use of the Joint Economic Committee Congress of the United States. Washington: Government Printing Office, 1973 (*a*)

———. "The Economics of the New Unemployment." *Public Interest* 33 (Fall 1973)· 3–42. (*b*)

[49] There is still a distortion in the search behavior for those who are permanently laid off or who can become eligible for unemployment insurance after quitting their previous job.

———. "Unemployment Compensation Adverse Incentives and Distributional Anomalies." *Nat. Tax J* 37, no 2 (June 1974) 231–44

———. "The Importance of Temporary Layoffs An Empirical Analysis." *Brookings Papers Econ Activity*, 1976, in press

Flemming, J S "Wage Rigidity and Employment Adjustment Alternative Micro-Foundations" Mimeographed 1974

Gordon, R. J. "The Welfare Cost of Higher Unemployment" *Brookings Papers Econ Activity*, no 1 (1973), pp 133-205

———. "The Microeconomic Foundations of Wage Ridigity" Mimeographed. January 1974.

Hall, R. E "Why Is the Unemployment Rate So High at Full Employment?" *Brookings Papers Econ Activity*, no 3 (1970), pp 369–402

——— "Turnover in the Labor Force" *Brookings Papers Econ Activity*, no 3 (1972) pp 709–56.

——— "The Process of Inflation in the Labor Market" *Brookings Papers Econ. Activity*, no 2 (1974), pp 343–409

Hamermesh, D S "A Disaggregative Econometric Model of Gross Changes in Employment" *Yale Econ Essays* (1969), pp 107–45

Hicks, J R *The Theory of Wages* London Macmillan, 1932

Holt, C C "How Can the Phillips Curve Be Moved to Reduce Both Inflation and Unemployment?" In *Microeconomic Foundations of Employment and Inflation Theory*, edited by E S Phelps New York Norton, 1970

Kaitz, H "Analyzing the Length of Spells of Unemployment" *Monthly Labor Rev* 93 (November 1970) 11–20

Lester, R A *The Economics of Unemployment Compensation* Princeton, N J Indus. Relations Sec, Princeton Univ, 1962

McCall, J J "Economics of Information and Job Search" *Q J E* 84 (February 1970) 113–25

Mortensen, D T "Job Search, the Duration of Unemployment and the Phillips Curve" *A E R* 60 (December 1970) 847–62

Oi, W Y "Labor as a Quasi-fixed Factor" *J P E* 70, no 6 (December 1962) 538–55

Parnes, H S, Nestel, G, and Andrisani, P *The Pre-Retirement Years* Longitudinal Study of the Labor Market Experience of Men, vol 3 (August 1972) Columbus. Center Human Resource Res, Ohio State Univ, 1972

Parsons, D O "Specific Human Capital An Application to Quit Rates and Layoff Rates" *J P E* 80, no 6 (November/December 1972) 1120–43.

Perry, G "Unemployment Flows in the U S Labor Market" *Brookings Papers Econ Activity*, no 2 (1972), pp 245–78

Phelps, E S, ed *Microeconomic Foundations of Employment and Inflation Theory* New York Norton, 1970

——— *Inflation Policy and Unemployment Theory* New York Norton, 1972.

Pigou, A. C *The Theory of Unemployment* London Macmillan, 1933

Stein, R L "Work History, Attitudes and Income of the Unemployed" *Monthly Labor Rev*, vol 86 (December 1963)

Stigler, G J. "The Economics of Information" *J P E* 69, no 3 (June 1961)· 312–25.

Tobin, J E "Inflation and Unemployment." *A E R*. 62 (March 1972). 1–18.

U S Department of Labor, Manpower Administration *The Pre-Retirement Years.* Longitudinal Study of the Labor Market Experience of Men, vol. 2 Manpower Research Monograph, no 15 Washington Government Printing Office, 1970.

U S. Department of Labor, Bureau of Labor Statistics *Layoff, Recall, and Worksharing Procedures*. Bulletin 1425–13. Washington Government Printing Office, 1972

———. *Employment and Earnings*, vol. 21, no 10 (April 1975)

Interfuel Substitution in Steam Electric Power Generation

Scott E. Atkinson

Federal Energy Administration

Robert Halvorsen

University of Washington

A translog normalized restricted profit function is used to study the characteristics of the production function for electric energy The results indicate that fuel choice in existing steam electric plants responds to changes in fuel prices The production function is also tested for separability of fuels from capital and labor, homotheticity, returns to scale, and embodied technical change

1. Introduction

Recent increases in energy prices and interruptions in energy supplies have led to increased interest in the characteristics of energy demand In particular, more information is desired concerning the extent to which market forces bring about optimal short-run reallocations of fossil fuels This paper examines the demand for fossil fuels by electric utilities, which accounted for 66 percent of coal, 8 percent of oil, and 19 percent of natural gas consumed as fuel in the United States in 1971 [1] The results include estimates of all own and cross elasticities of demand and substitution for fuels The paper also investigates other aspects of the production process, including separability of fuels from other inputs, homotheticity, returns to scale, and technical change

Research for this paper was supported by the Office of Energy Systems, Federal Energy Administration We are grateful to Ernst R Berndt and Lawrence J Lau for comments on an earlier version of this paper and are especially grateful to David Nissen for his advice and encouragement throughout the course of our research We retain responsibility for any remaining errors Views expressed in the paper are ours and do not reflect in any way policies of the Federal Energy Administration

[1] U S Bureau of the Census (1973, p 657)

[*Journal of Political Economy*, 1976, vol 84 no 5]
© 1976 by The University of Chicago All rights reserved

There have been a large number of econometric studies of steam electric power generation (see Nordin 1947; Lomax 1952; Iulo 1961; Komiya 1962, Nerlove 1963, Barzel 1963, 1964; Dhrymes and Kurz 1964; Ling 1964; McFadden 1964, Galatin 1968, Belinfante 1969, Cowing 1970; Fuss 1970, 1971, Lawrence 1972, Greene 1974, Joskow and Mishkin 1974) However, previous studies suffer from several shortcomings. First, all types of fuel have been aggregated into a single fuel input, precluding the possibility of estimating elasticities of substitution between fuels.[2] The validity of using a fuel aggregator has not previously been tested Second, production possibilities have been specified by restrictive functional forms, imposing severe a priori constraints on elasticities of substitution as well as other characteristics of the production process.[3] Finally, previous studies have generally assumed a putty-clay technology, again placing unduly severe restrictions on factor substitution [4]

The present study directly addresses these issues Coal, oil, and gas are treated as different inputs Instead of assuming a consistent fuel aggregator, we test for its existence To avoid imposing a priori restrictions on the characteristics of production, a transcendental logarithmic profit function is estimated Duality theory allows characteristics of the production function to be deduced from the estimated profit function. Finally, a putty-clay technology is not assumed for fuel substitution [5]

As discussed in later sections, our model does not incorporate all of the considerations affecting production decisions in electric power generation Also, the quality of the data for some variables is quite poor As a result, not all of the conditions implied by economic theory are satisfied by the estimated equations Therefore, while the study illustrates the application of a promising new methodology for research on production processes, our specific findings have to be considered as only tentative

Given this caveat, our results indicate that (1) substantial ex post substitution exists between fuels, (2) conditions for consistent aggregation of fuels are not in general satisfied, (3) the production function is predominately homothetic, (4) electric power production is not characterized by strongly increasing returns to scale, and (5) shifts in production functions due to embodied technical progress have been slight

Section 2 discusses the transcendental logarithmic profit function and the hypotheses to be tested Estimation of the profit function and factor demand functions is discussed in Section 3 Estimated elasticities of demand and substitution and results of the tests of hypotheses are summarized in Section 4. Section 5 contains concluding comments

---

[2] Exceptions are Lawrence (1972) and Joskow and Mishkin (1974) Lawrence used a branched CES production function to analyze interfuel substitution separately from capital and labor but did not test for the validity of this approach Joskow and Mishkin assumed that there are a limited number of fixed-coefficient technologies available to utilities and used a logit model

[3] Fuss (1971) is an exception

[4] Again, Fuss (1971) is an exception He tested putty-semiputty, putty-clay, and clay-clay hypotheses.

[5] However, capital and labor are assumed to be fixed inputs ex post.

2. The Model

The quantity of electric energy produced by a plant is assumed to be a function of variable and fixed inputs,

$$\begin{aligned} E &= F(X, Z) \\ &= F(X_1, \quad , X_m, Z_1, \quad , Z_n), \end{aligned}$$

where E is electric energy and X and Z are vectors of variable and fixed inputs, respectively Rather than estimate the production function directly, we estimate a normalized restricted profit function from which the properties of the underlying production function can be derived using duality theory

Normalized restricted profit is defined as

$$F(X, Z) - \sum_{i=1}^{m} P_i X_i,$$

where P_i is the price of input i divided by the price of output Let

$$\Pi = F(X^*, Z) - \sum_{i=1}^{m} P_i X_i^*,$$

where the asterisk indicates that the quantities of X are chosen to maximize normalized restricted profit Thus,

$$\begin{aligned} \Pi &= \Pi(P, Z) \\ &= \Pi(P_1, \quad , P_m, Z_1, \quad , Z_n) \end{aligned}$$

The specification of the restricted profit function is chosen to be

$$\Pi = \Pi(P_1, P_2, P_3, Z_1, Z_2, Z_3),$$

where P_1 = the price of coal, P_2 = the price of oil, P_3 = the price of gas, Z_1 = capital service input, Z_2 = labor input, Z_3 = the vintage of the capital, and all prices are normalized by dividing by the price of output

Estimation of a restricted profit function with capital as a fixed input avoids the necessity of modeling the effect of rate-of-return regulation on input choice [6] Labor is treated as a fixed input because adequate data are not available on labor prices The assumption that capital and labor are fixed inputs in existing plants is consistent with previous studies which assumed a putty-clay technology.[7] The use of a restricted profit function would be appropriate even if these inputs were not fixed.[8] However, our assumption that the arguments of the profit function are exogenous would not be satisfied and parameter estimates would not be consistent.[9]

---

[6] It has been shown that profit maximization subject to a rate-of-return constraint results in greater use of capital than would result from cost minimization (see Averch and Johnson 1962, Baumol and Klevorick 1970, Courville 1974, and Spann 1974).

The restricted profit function provides a convenient model for analyzing the effect of price changes on fuel choice. However, it does not explicitly reflect two nonprice influences on fuel choice. The first is the effect of environmental restrictions that exist in some areas. The second is the effect of changes in the temporal pattern of demand on fuel usage in plants which utilize different fuels for base-load and peak generation. These factors may have a significant effect on fuel choice, but data restrictions prevented their being incorporated into the model.

A number of flexible functional forms of the normalized restricted profit function could be estimated. The one used here is the transcendental logarithmic (translog):

$$\begin{aligned}\ln \Pi = {} & \alpha_0 + \sum_{i=1}^{3} \alpha_i \ln P_i + \frac{1}{2}\sum_{i=1}^{3}\sum_{h=1}^{3} \gamma_{ih} \ln P_i \ln P_h \\ & + \sum_{i=1}^{3}\sum_{j=1}^{3} \delta_{ij} \ln P_i \ln Z_j \qquad (1) \\ & + \sum_{j=1}^{3} \beta_j \ln Z_j + \frac{1}{2}\sum_{j=1}^{3}\sum_{k=1}^{3} \phi_{jk} \ln Z_j \ln Z_k,\end{aligned}$$

where $\gamma_{ih} = \gamma_{hi}$ and $\phi_{jk} = \phi_{kj}$ (see Lau 1969, and Christensen, Jorgenson, and Lau 1971, 1973).

By Hotelling's lemma,

$$\partial \Pi / \partial P_i = -X_i^*,$$

where X_i^* is the profit-maximizing amount of input i (see Diewert 1974a, pp. 42–45). Using this lemma, we find that logarithmic differentiation of the restricted profit function yields demand functions for each input:

$$\begin{aligned} -\frac{\partial \ln \Pi}{\partial \ln P_i} &= -\frac{\partial \Pi}{\partial P_i}\frac{P_i}{\Pi} = \frac{P_i X_i^*}{\Pi} = M_i \\ &= -\left(\alpha_i + \sum_{h=1}^{3} \gamma_{ih} \ln P_h + \sum_{j=1}^{3} \delta_{ij} \ln Z_j\right), \qquad i = 1, 2, 3, \end{aligned} \qquad (2)$$

where M_i is the ratio of expenditures on input i to restricted profits.

---

[7] Fuss concluded that the putty-clay hypothesis could not be rejected for a model incorporating capital, labor, and a fuel aggregate. Barzel (1964) and Dhrymes and Kurz (1964) present evidence that labor is largely fixed in existing plants.

[8] If capital and labor are not fixed, the restricted profit function is interpretable as the first step in a two-step maximization, the second step involving maximization with respect to capital and labor (see Lau 1974a, pp. 9–10).

[9] The normalized fuel price variables can be assumed to be exogenous because fuel prices are determined in relatively competitive markets and the price of output is determined by regulatory authorities.

The own- and cross-price elasticities of demand for fuels are readily derived from the demand functions The own-price elasticity is defined as

$$E_{ii} = \frac{\partial X_i^*}{\partial P_i}\,\frac{P_i}{X_i^*}.$$

By Hotelling's lemma,

$$X_i^* = -\frac{\partial \Pi}{\partial P_i} \quad \text{and} \quad \frac{\partial X_i^*}{\partial P_i} = -\frac{\partial^2 \Pi}{\partial P_i^2}$$

Therefore,

$$E_{ii} = \frac{P_i \cdot \partial^2 \Pi/\partial P_i^2}{\partial \Pi/\partial P_i} = \frac{-M_i^2 - M_i - \gamma_{ii}}{M_i} \tag{3}$$

Similarly, the cross-price elasticity is equal to

$$E_{ih} = \frac{\partial X_i^*}{\partial P_h}\,\frac{P_h}{X_i^*} = \frac{P_h \;\; \partial^2 \Pi/\partial P_i \partial P_h}{\partial \Pi/\partial P_i} = \frac{-M_i M_h - \gamma_{ih}}{M_i} \tag{4}$$

Partial elasticities of substitution are defined as[10]

$$\sigma_{ii} = \frac{1}{M_i} E_{ii} = \frac{-M_i^2 - M_i - \gamma_{ii}}{M_i^2}, \tag{5a}$$

$$\sigma_{ih} = \frac{1}{M_h} E_{ih} = \frac{-M_i M_h - \gamma_{ih}}{M_i M_h} \tag{5b}$$

The partial elasticities of substitution are normalized price elasticities, where the normalization is chosen such that the elasticities of substitution are invariant to the ordering of the factors Therefore, $\sigma_{ih} = \sigma_{hi}$ although, in general, $E_{ih} \neq E_{hi}$

As indicated above, most previous studies have assumed that the production technology can be specified in terms of capital, labor, and a fuel aggregate[11] This approach is valid if fuels are weakly separable from capital and labor[12] The model estimated here makes it possible to test the separability assumption explicitly

---

[10] The partial elasticities defined here are analogous but not identical to Allen elasticities of substitution They differ in two respects quantities rather than prices of fixed factors are held constant, and the price rather than the quantity of output is held constant

[11] See Nerlove (1963), Barzel (1964), Dhrymes and Kurz (1964), McFadden (1964), and Fuss (1971)

[12] In the literature the general procedure is that, given fuel consumption and cost data, the quantity index is a weighted quantity sum (e g , total Btu) and the price index is average cost The use of such a procedure in defining costs is called "strong additive price aggregation" by Blackorby, Primont, and Russell (1975) and requires that the production function be groupwise homogeneous and groupwise separable In this case, the restricted profit function is groupwise homogeneous and groupwise separable in the dual prices David Nissen brought this point to our attention

The production function is weakly separable with respect to fuels if

$$E = F[f(C, O, G), K, L],$$

where E is output, f is a consistent aggregator, and C, O, G, K, and L are coal, oil, gas, capital, and labor, respectively. A necessary and sufficient condition for a homogeneous (homothetic) production function to be weakly separable with respect to C, O, and G is that the restricted profit function be homogeneous (homothetic) and weakly separable with respect to the prices of coal, oil, and gas, P_C, P_O, and P_G (see Lau 1969, pp. 19–20). A necessary and sufficient condition for the production function to be groupwise homogeneous (homothetic) and groupwise weakly separable with respect to C, O, and G is that the restricted profit function be groupwise homogeneous (homothetic) and groupwise weakly separable with respect to P_C, P_O, and P_G.[13]

The restricted profit function will be separable in the prices of fuels r and s from the quantity of fixed input t if

$$\delta_{rt} = \delta_{st} = 0, \qquad r, s = 1, 2, 3, \qquad t = 1, 2, \tag{6}$$

or if

$$\frac{\alpha_r}{\alpha_s} = \frac{\gamma_{rh}}{\gamma_{sh}} = \frac{\delta_{rj}}{\delta_{sj}}, \qquad r, s, h = 1, 2, 3, \qquad j = 1, 2, 3 \tag{7}$$

(see Berndt and Christensen 1973*b*). Restrictions (6) and (7) will be referred to as the linear and nonlinear separability restrictions, respectively. In the case of nonlinear separability, separability from either capital or labor implies separability from both and from capital, labor, and vintage.

It should be noted that the separability restrictions are equivalent to equality restrictions on the elasticities of substitution (see Berndt and Christensen 1973*a*). The linear separability restrictions imply that the elasticities of substitution between fuels and other inputs are equal to 1, $\sigma_{rt} = \sigma_{st} = 1$, where r and s are fuel inputs and t is capital or labor. The nonlinear restrictions imply that the elasticities are equal for all fuel inputs but not equal to 1, $\sigma_{rt} = \sigma_{st} \neq 1$.

The production function is homothetic if and only if the restricted profit function satisfies the equation

$$-\sum_{i=1}^{3} P_i \frac{\partial \Pi}{\partial P_i} + \sum_{j=1}^{2} \frac{\partial \Pi}{\partial Z_j} Z_j = J\left(\Pi - \sum_{i=1}^{3} P_i \frac{\partial \Pi}{\partial P_i}\right), \tag{8}$$

where J is an arbitrary nonnegative and differentiable function defined on the nonnegative real line with $J(0) = 0$ (see Lau 1974*b*, pp. 4–7).

---

[13] Diewert (1974*b*) uses the terminology "homothetically weakly separable" to mean groupwise homothetically weakly separable.

Necessary and sufficient restrictions for the translog function to satisfy equation (8) are

$$\lambda \sum_{i=1}^{3} \gamma_{ir} + \sum_{j=1}^{2} \delta_{jr} = \sigma\alpha_r, \qquad r = 1, 2, 3, \tag{9a}$$

$$\lambda \sum_{i=1}^{3} \delta_{is} + \sum_{j=1}^{2} \phi_{js} = \sigma\beta_s, \qquad s = 1, 2, \tag{9b}$$

where

$$\sigma \equiv J'\left(1 - \sum_{i=1}^{3} \frac{\partial \ln \Pi}{\partial \ln P_i}\right) - \frac{J}{\Pi} \qquad \text{and} \qquad \lambda \equiv J' - 1$$

at the point of approximation[14]

Homotheticity of the production function is of interest for its own sake and will be tested for whether or not the restricted profit function is separable The production function will also be tested for the more restrictive condition of homogeneity The production function is homogeneous of degree k, a scalar constant, if and only if the restricted profit function satisfies the equation

$$-\sum_{i=1}^{3} P_i \frac{\partial \Pi}{\partial P_i} + \sum_{j=1}^{2} \frac{\partial \Pi}{\partial Z_j} Z_j = k\left(\Pi - \sum_{i=1}^{3} P_i \frac{\partial \Pi}{\partial P_i}\right), \tag{10}$$

which is a specialization of equation (8)

When the restricted profit function satisfies (10), J' is a constant everywhere, not just at the point of expansion, and the value of J' is equal to the degree of returns to scale, k Since

$$J = J'\Pi\left(1 - \sum_{i=1}^{3} \frac{\partial \ln \Pi}{\partial \ln P_i}\right),$$

σ in equations (9a) and (9b) is equal to zero Thus, homogeneity can be tested by adding one additional restriction to those for homotheticity, and the degree of returns to scale can be determined by calculating $J' = \lambda + 1$

The production function is also tested for groupwise homotheticity and homogeneity The production function is homothetic in the variable inputs if and only if the restricted profit function satisfies

$$\sum_{i=1}^{3} \gamma_{ik} = \sigma\alpha_k, \qquad k = 1, 2, 3 \tag{11}$$

The production function is homogeneous of degree k in the variable inputs if and only if the restricted profit function is homogeneous of degree $-k/(1-k)$ in normalized prices (see Lau 1969, pp. 101–2). This will be true if and only if

---

[14] The point of approximation is the geometric mean of the sample

$$\sum_{i=1}^{3} \gamma_{ih} = 0, \qquad h = 1, 2, 3 \tag{12a}$$

$$\sum_{i=1}^{3} \delta_{ij} = 0, \qquad j = 1,2 \tag{12b}$$

The normalized restricted profit function can also be used to examine the characteristics of technical change in steam electric generation. If the production function is homothetic, Hicksian neutrality of the restricted profit function implies Hicksian neutrality of the production function (see Lau 1969, pp 45–49) If the production function is not homothetic, Hicksian neutrality of the restricted profit function implies that technical change is "output augmenting" and "factor saving" at equal rates The restricted profit function will be Hicksian neutral if

$$\delta_{i3} = 0, \qquad i = 1, 2, 3, \tag{13a}$$

$$\phi_{j3} = 0, \qquad j = 1, 2 \tag{13b}$$

The absence of technical change implies restrictions (13a) and (13b) plus

$$\phi_{33} = 0, \tag{14a}$$

$$\beta_3 = 0 \tag{14b}$$

3. Estimation

The set of equations to be estimated consists of the restricted profit equation, (1), and the factor demand equations, (2) A classical additive disturbance term is included in each equation to reflect errors in profit-maximizing behavior

Because cross-equation restrictions are imposed in performing tests of hypotheses, the equations are estimated simultaneously The off-diagonal elements of the residual covariance matrix can be expected to be nonzero, and therefore iterative Zellner-efficient estimation is used [15] The estimates will be consistent under the assumption that all explanatory variables are exogenous

A normalized restricted profit function is well behaved only if it is strictly decreasing and convex in the normalized prices and strictly increasing in fixed inputs Lau (1976) has proposed statistical tests for monotonicity and convexity, but technical problems in imposing the required nonnegativity constraints have prevented their implementation to date. However, some indication of the extent to which the fitted restricted profit function is well behaved can be obtained by examining the regression results for each observation

---

[15] Iterative Zellner-efficient estimation is equivalent to maximum-likelihood estimation of a reduced form (see Zellner 1962, Oberhofer and Kmenta 1974).

The condition that the restricted profit function be strictly decreasing in normalized prices is checked by examining the fitted expenditure ratios for positivity, since

$$M_i = -\frac{\partial \Pi}{\partial P_i}\frac{P_i}{\Pi}$$

and P_i and Π are positive for all observations in the sample The condition that the restricted profit function be strictly increasing in fixed inputs is checked by calculating

$$\frac{\partial \Pi}{\partial Z_j} = \frac{\Pi}{Z_j}\left(\sum_{i=1}^{3} \delta_{ij} \ln P_i + \beta_j + \sum_{k=1}^{3} \phi_{jk} \ln Z_k\right), \qquad j = 1, 2,$$

for each observation

Convexity in prices requires that the Hessian of the restricted profit function be positive definite for each observation This will be true if and only if all principal minors are positive Convexity is checked for each observation by determining the signs of the estimated principal minors Since it is not determined if any negative principal minors are statistically significant, this procedure does not constitute a statistical test of convexity

The equations are estimated for individual plants using Federal Power Commission data for 1972 [16] The prices of fuels are calculated by dividing expenditures on each type of fuel by quantity purchased Therefore, plants which used only one type of fuel in 1972 could not be included in the sample [17] The number of plants using all three fuels is too small for estimation of the system of profit and demand equations Therefore, separate sets of equations are estimated for each pair of fuels coal-gas, coal-oil, and oil-gas

Each system of equations consists of a normalized restricted profit function and two fuel demand equations Derivation of the demand equations from the profit equation implies 12 equality restrictions relating the coefficients of the demand equation to those of the profit equation In order to test whether or not the loss of fit from imposing the equality restrictions is significant, the equations are estimated with and without the restrictions imposed The results are compared by computing −2 log λ, where λ is the ratio of the maximum value of the likelihood function for the restricted equations to the maximum value of the likelihood function for the unrestricted equations This statistic, computed to test all restrictions, is distributed asymptotically as chi-squared with degrees of freedom equal to the number of restrictions[18]

---

[16] Data sources are discussed in Appendix B

[17] Plants using only trace amounts of one of the fuels were also omitted from the sample The direction of the bias, if any, resulting from the omission of plants using only one type of fuel is not clear Plants using only one fuel may have gone to the extreme of substituting entirely away from other fuels or may not be able to use more than one fuel Estimated elasticities would be biased downward in the former case and upward in the latter case

Given the equality restrictions, it is neither possible nor necessary to test for the symmetry conditions, $\gamma_{ih} = \gamma_{hi}$ and $\phi_{jk} = \phi_{kj}$, because symmetry does not impose any additional effective constraints on the translog form For example, given the equality restrictions, if symmetry is not imposed, each coefficient γ_{ih} is replaced by $\frac{1}{2}$ $(\gamma_{ih} + \gamma_{hi})$. Because the two nonsymmetric coefficients always appear together, they cannot be separately identified, and no loss of fit is entailed in substituting for them the single coefficient γ_{ih}

4. Empirical Results

Parameter estimates and asymptotic standard errors for the set of restricted profit and factor demand equations for each pair of fuels are presented in table 1 The number of observations is 31 for the coal-gas sample, 17 for the coal-oil sample, and 60 for the oil-gas sample The table also shows the value of $\bar{R}^2$ for each equation, where $\bar{R}^2$ is computed as 1 minus the ratio of the residual sum of squares to the total sum of squares and then adjusted for the number of parameters estimated The calculated $\bar{R}^2$s are quite high for cross-section equations fitted with microdata However, the standard errors for some parameters are also quite large

The condition that the restricted profit function be decreasing in the prices of variable inputs implies that the fitted expenditure ratios should be positive All but two of the 62 fitted ratios are positive for the coal-gas sample, and all are positive for the oil-gas sample The coal-oil sample performs somewhat less well, with six of the 34 fitted ratios negative

A second condition for a well-behaved restricted profit function is that the partial derivatives of restricted profit with respect to fixed inputs be positive The calculated values of the derivative with respect to capital are all positive for the three samples The results for the derivative of restricted profit with respect to labor are less satisfactory Sixteen of the calculated derivatives are negative for the coal-gas and oil-gas samples, and eight are negative for the coal-oil sample This may be due to the poor quality of the data on labor input, which is measured as average number of employees per year with no adjustment for the number of hours worked per employee

---

[18] Berndt and Savin (1974) examine alternative test statistics in the multivariate linear regression model and find that the value of the likelihood-ratio test statistic lies between the values of the Wald and Lagrange multiplier test statistics

TABLE 1

ESTIMATED PARAMETERS

PARAMETER	COAL-GAS SAMPLE		COAL-OIL SAMPLE		OIL-GAS SAMPLE	
	Estimate	SE	Estimate	SE	Estimate	SE
α_0	62 765	27 655	−59 843	41 160	−11 218	23.869
α_1	0 658	0 398	1 291	0 758		
α_2			−6 049	1 668	1 405	0 653
α_3	−0 422	0 289			0 351	0 452
γ_{11}	−0 263	0 048	−0 024	0 082		
γ_{22}			0 064	0 163	0 073	0 061
γ_{33}	0 023	0 030			−0 324	0 052
γ_{13}	−0 046	0 028				
γ_{12}			−0 309	0 058		
γ_{23}					−0 167	0 041
δ_{11}	−0 016	0 033	−0 020	0 045		
δ_{12}	−0 008	0 038	−0 025	0 060		
δ_{13}	0 074	0 047	0 044	0 056		
δ_{21}			−0 461	0 154	0 085	0.042
δ_{22}			0 551	0 135	−0 144	0 062
δ_{23}			1 028	0 212	−0 055	0 065
δ_{31}	0 003	0 024			−0 011	0 029
δ_{32}	0 021	0 027			0 083	0 041
δ_{33}	0 058	0 034			0 044	0 043
β_1	3 405	3 671	−4 378	2 293	−4 452	3 116
β_2	−5 428	4 842	15 791	7 076	5 387	4 756
β_3	−15 736	6 508	3 410	3 116	5 427	4 970
ϕ_{11}	−0 061	0 318	0 295	0 186	−0 573	0.298
ϕ_{22}	0 107	0 426	−1 591	0 636	−0 571	0 486
ϕ_{33}	1 034	0 743	−0 563	0 281	−0 358	0 642
ϕ_{13}	−0 866	0 483	−0 150	0 257	0 606	0 364
ϕ_{12}	0 025	0 347	0 469	0 235	0 586	0 331
ϕ_{23}	1 449	0 613	−0 411	0 283	−0 655	0 503
			Equation R^2			
Profit	0 95		0 96		0 89	
Coal	0 75		0 69			
Oil			0 44		0 17	
Gas	0 37				0 40	

Convexity of the restricted profit function is checked by determining whether the values of the principal minors of the estimated Hessian are positive Negative values were found for nine observations in the coal-gas sample, 10 in the coal-oil sample, and 51 in the oil-gas sample[19] Thus, it appears that the restricted profit function fails to satisfy the convexity condition for a substantial number of observations Since convexity is implied by the assumption of profit maximization, this is a serious weakness in the performance of the model Tests of the equality restrictions relating parameters in the profit and demand equations also cast doubt on the performance of the model These restrictions, which are also implied by profit maximization, were rejected for the coal-gas and coal-oil samples.[20]

---

[19] The total number of principal minors is 62 for the coal-gas sample, 34 for the coal-oil sample, and 120 for the oil-gas sample

There are several possible reasons for the problems with convexity and equality As discussed above, the model does not explicitly incorporate the effects on fuel choice of environmental restrictions or of changes in plant load factors Also, there are serious weaknesses in the data used In particular, output price for each plant had to be approximated by dividing the firm's total revenue minus transmission and distribution expenses by total sales This measure of the value of output can be expected to result in an overstatement of normalized restricted profit for some plants and an understatement for others Last, the cross-section data used may include observations that are too far from the point of approximation of the translog function to the true profit function These shortcomings should be kept in mind in evaluating the results presented below

Estimated elasticities of demand and substitution, evaluated at the means of the data, are shown in table 2 Own-price elasticities should be negative and the cross-price elasticities should be positive if fuels are substitutes Eleven of the 12 price elasticities in table 2 have the appropriate sign, of which 10 are significant at the 5 percent level The own-price elasticity for gas in the oil-gas sample has an incorrect sign but is insignificant Eight of the nine elasticities of substitution have the correct sign, and seven of these are significant The own elasticity for gas in the oil-gas sample again has an incorrect sign but is insignificant

Estimated price elasticities and elasticities of substitution for individual plants are presented in Tables A1 and A2 Plants are listed in increasing order of size For brevity, estimated elasticities are shown for only odd-numbered plants The values of the estimates for individual plants are similar to the elasticities evaluated at the means of the data There are few extreme values The total number of estimated price elasticities is 432, and all but 69 have the appropriate signs There are a total of 324 elasticities of substitution, of which all but 68 have the correct signs The incorrect signs for some elasticities reflect the apparent failure of the profit function to be convex for all observations

As indicated by the size of the standard errors shown in table 2, some of the estimates of the price elasticities and elasticities of substitution have large confidence intervals However, their general magnitudes are consistent with the hypothesis that fuels are substantially but not perfectly substitutable in existing steam electric plants

The results of statistical tests of joint hypotheses concerning other characteristics of the production function are summarized in table 3. An upper bound for the overall significance level for the group of tests

---

[20] The test statistics were 34 3 for coal-gas, 257.1 for coal-oil, and 18 4 for oil-gas The critical value at the 01 level is 26 2

TABLE 2

ESTIMATED ELASTICITIES OF DEMAND AND SUBSTITUTION[a]

Coal-Gas	Coal-Oil	Oil-Gas
	Price Elasticities of Demand[b]	
$E_{CC} = -0.43^*$ (0.16)	$E_{CC} = -1.15^*$ (0.33)	$E_{OO} = -1.60^*$ (0.36)
$E_{CG} = 0.09$ (0.47)	$E_{CO} = 0.99^*$ (0.24)	$E_{OG} = 0.76^*$ (0.24)
$E_{GC} = 0.45^*$ (0.10)	$E_{OC} = 1.01^*$ (0.24)	$E_{GO} = 0.58^*$ (0.18)
$E_{GG} = -1.43^*$ (0.48)	$E_{OO} = -1.50^*$ (0.67)	$E_{GG} = 0.21$ (0.23)
	Elasticities of Substitution	
$\sigma_{CC} = -1.43^*$ (0.54)	$\sigma_{CC} = -4.63^*$ (1.32)	$\sigma_{OO} = -9.46^*$ (2.14)
$\sigma_{GC} = 1.50$ (1.56)	$\sigma_{CO} = 4.06^*$ (0.96)	$\sigma_{OG} = 3.36^*$ (1.06)
$\sigma_{GG} = -22.81^*$ (7.66)	$\sigma_{OO} = -6.14^*$ (2.72)	$\sigma_{GG} = 0.92$ (1.02)

[a] Evaluated at the means of the data. Figures in parentheses are asymptotic standard errors. The subscripts *C*, *O*, and *G* stand for coal, oil, and gas, respectively.

[b] For cross-elasticities, the quantity of the fuel indicated by the first subscript is a function of the price of the fuel indicated by the second subscript.

* Significant at .05 level.

TABLE 3

TESTS OF HYPOTHESES

	NUMBER OF RESTRICTIONS	χ^2 CRITICAL VALUE*	TEST STATISTICS		
			Coal-Gas	Coal-Oil	Oil-Gas
Linear separability of fuels					
From capital and labor	4	13.3	3.2†	8.8†	14.1
From capital, labor, and vintage	6	16.8	17.8	29.5	22.3
Nonlinear separability of fuels	5	15.1	8.7†	24.6	24.9
Homotheticity:					
All inputs	2	9.2	16.2	6.9†	6.1†
Fuels	1	6.6	14.1	4.7†	5.9†
Homogeneity					
All inputs	3	11.3	26.5	8.8†	6.5†
Fuels	4	13.3	40.3	11.7†	60.4
Absence of technical change	6	16.8	16.0†	29.5	4.2†
Hicksian neutrality	4	13.3	15.3	19.4	3.4†

* The level of significance for each test is .01.

† The null hypothesis is accepted.

is the sum of the individual significance levels. We assign an overall significance level of .09 and allocate a significance level of .01 to each of the nine χ^2 tests of our joint hypotheses.[21] In performing the tests, we imposed the equality restrictions relating parameters in the profit and demand equations as maintained hypotheses.

Linear separability of fuel prices from capital, labor, and vintage is rejected for all samples. Linear separability of fuel prices from capital and labor is rejected for the oil-gas sample but accepted for the coal-gas and coal-oil samples Nonlinear separability of fuel prices from any, and therefore all, fixed inputs is rejected for the coal-oil and oil-gas samples but accepted for the coal-gas sample. Overall homotheticity as well as groupwise homotheticity in fuels is rejected for the coal-gas sample but accepted for the coal-oil and oil-gas samples Therefore, a consistent fuel aggregator is accepted for only the coal-oil sample

As shown in table 3, overall homogeneity is accepted for the oil-gas and coal-oil samples [22] Therefore, the degree of returns to scale for these samples can be calculated as $\lambda + 1$ For the oil-gas sample, the computed degree of homogeneity is 1 10 and the standard error of λ is 0 04. Thus, returns to scale are significantly greater than unity for this sample, but the extent of increasing returns is not very great [23] The 95 percent confidence interval is 1 02–1 18 The degree of returns to scale in the coal-oil sample is 0 95 and the standard error of λ is 0 20 [24] Therefore, the estimated degree of returns to scale is not significantly greater than unity. However, the estimate is not very precise the 95 percent confidence interval is 0 56–1 34

Last, the restricted profit function is used to test for the existence and neutrality of embodied technical change. For the coal-oil sample, both absence and Hicksian neutrality of technical change are rejected Absence of technical change is accepted for the coal-gas and oil-gas samples, but Hicksian neutrality of the profit function is accepted only for the oil-gas sample

5. Concluding Comments

Estimation of a translog normalized restricted profit function makes it possible to calculate the extent of interfuel substitution in conventional steam electric power generation without imposing a priori constraints on the values of the elasticities. In addition, the use of a flexible functional form makes it possible to test a number of hypotheses concerning other characteristics of the production process. The results indicate that steam electric power generation is characterized by substantial ex post interfuel substitution, homotheticity, absence of strong scale economies, and little embodied technical change.

---

[21] An alternative testing procedure would have been to perform Bonferroni *t*-tests on the joint hypotheses However, Christensen (1973) has shown that the power of this test procedure is likely to be quite low.

[22] This result conflicts with the conclusion of Nerlove (1963) that the production function may be homothetic but is not homogeneous However, Nerlove's results were for firms rather than plants

[23] The extent of increasing returns is defined as the degree of homogeneity minus 1

[24] For the coal-oil sample, the degree of returns to scale for fuel inputs can also be computed and is equal to 0.73.

These findings contradict some widely held beliefs about production of electric power and have potentially important implications for public policy. However, due to weaknesses in the data and in the performance of the model, further work is required to confirm the results. We believe that the translog normalized restricted profit function offers a fruitful approach for further research in this area.

Appendix A

TABLE A1

ESTIMATED PRICE ELASTICITIES

1. COAL-GAS SAMPLE

Plant	E_{CC}	E_{CG}	E_{GC}	E_{GG}
1	−0.67	−0.05	−0.10	−1.31
3	0.16	0.15	0.32	−1.35
5	−0.94	−0.05	−0.17	−1.31
7	−0.26	0.08	0.25	−1.35
9	0.69	0.28	0.95	−1.60
11	−0.84	0.001	0.01	−1.33
13	−0.77	0.03	0.15	−1.37
15	−0.02	0.17	1.07	−1.69
17	−0.71	0.07	0.49	−1.50
19	−0.58	0.09	0.77	−1.60
21	−0.35	0.15	2.30	−2.32
23	−0.62	0.09	0.83	−1.64
25	−0.88	0.01	0.04	−1.35
27	−0.32	0.14	1.20	−1.78
29	−0.48	0.12	1.62	−2.00
31	7.43	1.51	−1.46	−0.25

2. COAL-OIL SAMPLE

Plant	E_{CC}	E_{CO}	E_{OC}	E_{OO}
1	−1.02	1.65	1.10	−1.50
3	−1.41	−0.87	−0.26	−2.58
5	−1.20	1.30	−1.72	−0.49
7	−1.04	0.96	0.22	−1.87
9	−1.25	1.01	−6.48	0.32
11	−1.18	0.91	1.08	−1.51
13	−0.28	9.89	−1.95	−0.44
15	−1.35	0.88	−2.86	−0.37
17	−1.39	0.50	1.12	−1.52

TABLE A1—*Continued*

3 OIL-GAS SAMPLE

Plant	E_{OO}	E_{OG}	E_{GO}	E_{GG}
1	−1.60	0.66	0.33	−0.36
3	−1.55	0.52	0.53	0.25
5	−1.55	0.14	0.10	−0.61
7	−1.80	1.34	0.59	0.11
9	−1.72	1.06	0.45	−0.18
11	−1.55	0.49	0.39	−0.08
13	−1.56	0.51	0.33	−0.27
15	−1.92	1.59	0.44	−0.29
17	−1.56	0.32	0.13	−0.83
19	−1.65	0.88	0.48	−0.06
21	−1.77	1.25	0.57	0.09
23	−2.00	1.87	0.70	0.30
25	−1.59	0.66	0.43	−0.10
27	−1.54	0.42	0.84	1.07
29	−1.55	0.26	0.31	−0.05
31	−1.54	0.57	1.89	3.11
33	−1.62	0.98	2.65	4.40
35	−1.98	1.84	0.73	0.38
37	−1.91	1.67	0.79	0.53
39	−1.62	0.80	0.51	0.05
41	−1.56	0.40	2.21	3.93
43	−1.66	0.93	0.54	0.08
45	−1.68	0.96	0.45	−0.16
47	−1.74	1.28	1.11	1.25
49	−1.89	1.46	0.39	−0.41
51	−1.65	0.91	0.54	0.08
53	−1.65	0.91	0.54	0.08
55	−1.55	0.60	1.05	1.35
57	−3.48	5.29	0.58	−0.10
59	−1.60	0.77	0.67	0.44

TABLE A2

ESTIMATED ELASTICITIES OF SUBSTITUTION

1 COAL-GAS SAMPLE

Plant	σ_{CC}	σ_{CG}	σ_{GG}
1	−1.79	−0.28	−7.61
3	0.84	1.67	−15.19
5	−1.94	−0.35	−8.91
7	−1.00	0.93	−14.88
9	4.84	6.64	−37.93
11	−1.91	0.01	−12.77
13	−1.88	0.36	−16.65
15	−0.11	4.88	−47.51
17	−1.83	1.27	−28.43
19	−1.69	2.21	−38.68
21	−1.24	8.16	−130.22
23	−1.74	2.31	−42.13
25	−1.93	0.09	−14.63
27	−1.16	4.35	−56.75
29	−1.52	5.16	−84.01
31	238.94	−46.85	7.61

TABLE A2—*Continued*

2 COAL-OIL SAMPLE

Plant	σ_{CC}	σ_{CO}	σ_{OO}
1	-6.23	6.75	-6.16
3	-3.06	-0.56	-1.57
5	-4.22	-6.03	2.27
7	-5.89	1.22	-2.37
9	-3.86	-20.02	-6.29
11	-4.37	3.98	-6.56
13	-8.91	-61.33	2.76
15	-3.31	-7.01	2.93
17	-3.15	2.54	-7.66

3 OIL-GAS SAMPLE

Plant	σ_{OO}	σ_{OG}	σ_{GG}
1	-9.57	1.97	-1.09
3	-6.89	2.37	1.16
5	-5.05	0.34	-1.50
7	-17.06	5.57	0.45
9	-13.96	3.65	-0.60
11	-7.14	1.81	-0.31
13	-7.70	1.64	-0.86
15	-21.94	5.03	-0.91
17	-7.59	0.65	-1.69
19	-11.35	3.27	-0.22
21	-15.85	5.14	0.37
23	-24.96	8.73	1.41
25	-8.92	2.39	-0.36
27	-5.24	2.87	7.32
29	-4.91	0.98	-0.19
31	-6.03	7.41	40.11
33	-10.10	16.52	14.22
35	-24.30	9.01	1.88
37	-21.21	8.83	2.81
39	-10.16	3.19	0.20
41	-4.39	6.22	60.44
43	-11.76	3.81	0.31
45	-12.58	3.33	-0.56
47	-14.78	9.42	9.22
49	-20.48	4.22	-1.18
51	-11.47	3.73	0.34
53	-11.41	3.71	0.33
55	-6.79	4.59	10.34
57	-116.36	19.26	-0.35
59	-9.27	3.87	2.20

Appendix B

Data Sources and Definitions

Unless noted otherwise, all data are for individual steam electric plants for 1972. The data are from Federal Power Commission (FPC) publications (1973*a*, 1973*b*, 1974).

Normalized fuel prices are calculated by dividing cost per 10^6 Btu by the price of output. Data on expenditure per 10^6 Btu for each type of fuel are from FPC

(1974). In order to reflect the effect of environmental restrictions on fuel costs, the price of coal is adjusted by adding the cost per 10^6 Btu of ash removal. Data on the cost of ash removal are for 1970 and are from FPC (1973*b*).

Data on the price of output are not available at the plant level Therefore, data from FPC (1973*a*) for the firm owning each plant are used To obtain a measure of the net value of plant output to the firm, output price is calculated by dividing total revenue minus transmission and distribution expenses by total sales

The average number of employees for each plant from FPC (1974) is used as the measure of labor input The quantity of capital is equal to installed generating capacity, which is measured by the manufacturer's maximum nameplate rating (FPC 1974) The vintage of the capital for each plant is calculated by weighting the vintage of each generating unit by its capacity (FPC 1974) Restricted profits are calculated as net generation (FPC 1974) minus normalized expenditures on fuels

The sample includes all plants beginning operation in 1946 or later for which the necessary data are available Data on fuel prices are available only for plants actually purchasing the fuels in question Therefore, the sample had to be restricted to plants purchasing more than one type of fuel There are 31 plants in the coal-gas sample, 17 in the coal-oil sample, and 60 in the oil-gas sample A list of the included plants is available from the authors on request

References

Averch, Harvey, and Johnson, Leland L "Behavior of the Firm under Regulatory Constraint " *A E R* 52, no 5 (December 1962) 1053–69

Barzel, Yoram "Productivity in the Electric Power Industry, 1929–1955 " *Rev Econ and Statis* 45, no 4 (November 1963) 395–408

——— "The Production Function and Technical Change in the Steam-Power Industry " *J P E* 72, no 2 (April 1964) 133–50

Baumol, William J , and Klevorick, Alvin K "Input Choices and Rate of Return Regulation An Overview of the Discussion " *Bell J Econ and Management Sci* 1, no 2 (Autumn 1970) 162–90

Belinfante, Alexander E "Technical Change in the Steam Electric Power Generating Industry " Ph D dissertation, Univ California, Berkeley, 1969

Berndt, Ernst, and Christensen, Laurits R "The Internal Structure of Functional Relationships Separability, Substitution, and Aggregation " *Rev Econ Studies* 40 (3), no 123 (July 1973) 403–10 (*a*)

——— "The Translog Function and the Substitution of Equipment, Structures, and Labor in U S Manufacturing 1929–1968 " *J Econ* 1, no 1 (March 1973) 81–113 (*b*)

Berndt, Ernst R , and Savin, Eugene N "Conflict among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model " Discussion Paper no 74–21 Vancouver Dept Econ , Univ British Columbia, October 1974.

Blackorby, C , Primont, D , and Russell, R R "Budgeting, Decentralization, and Aggregation " *Ann Econ and Soc Measurement* 4, no 1 (Winter 1975). 23–44.

Christensen, Laurits R "Simultaneous Statistical Inference in the Normal Multiple Linear Regression Model " *J American Statis Assoc* 68, no. 342 (June 1973) 457–61

Christensen, Laurits R., Jorgenson, Dale W , and Lau, Lawrence J "Conjugate Duality and the Transcendental Logarithmic Production Function." *Econometrica* 39, no. 4 (July 1971) 255–56.

——— "Transcendental Logarithmic Production Frontiers" *Rev Econ and Statis.* 55, no. 1 (February 1973) 28–45

Courville, Leon. "Regulation and Efficiency in the Electric Utility Industry" *Bell J. Econ and Management Sci* 5, no 1 (Spring 1974) 53–74

Cowing, Thomas G "Technical Change in Steam-Electric Generation An Engineering Approach" Ph D dissertation, Univ California, Berkeley, 1970

Dhrymes, Phoebus J., and Kurz, Mordecai "Technology and Scale in Electricity Generation." *Econometrica* 32, no 3 (July 1964) 287–315

Diewert, W E "Applications of Duality Theory" In *Frontiers in Quantitative Economics*, edited by M Intriligator and D Kendrick Vol 2 Amsterdam North-Holland, 1974 (*a*)

——— "Homogeneous Weak Separability and Exact Index Numbers" Technical Report no 122 Stanford, Calif Inst Math Studies Soc Sci, Stanford Univ, January 1974 (*b*)

Fuss, Melvyn A "The Time Structure of Technology An Empirical Analysis of the Putty-Clay Hypothesis" Ph D dissertation, Univ California, Berkeley, 1970

——— "Factor Substitution in Electricity Generation A Test of the Putty-Clay Hypothesis" Harvard Institute of Economic Research, Discussion Paper no 185 Cambridge, Mass Harvard Univ, 1971

Galatin, Malcolm *Economies of Scale and Technological Change in Thermal Power Generation.* Amsterdam North-Holland, 1968

Greene, William H "Factor Substitution and Returns to Scale in Electrical Supply" Unpublished paper Madison Univ Wisconsin, 1974

Iulo, William *Electrical Utilities—Costs and Performance* Pullman Washington State Univ Press, 1961

Joskow, Paul L, and Mishkin, Frederic S "Electric Utility Fuel Choice in the United States" Mimeographed Cambridge, Mass Massachusetts Inst Tech, October 1974

Komiya, Ryutaro "Technological Progress and the Production Function in the United States Steam Power Industry" *Rev Econ and Statis* 44, no 2 (May 1962) 156–66

Lau, Lawrence J "Some Applications of Profit Functions" Memorandum no 86A and 86B Stanford, Calif Res Center Econ Growth, Stanford Univ., November 1969

——— "A Characterization of the Normalized Restricted Profit Function." Technical Report no 134 Stanford, Calif Inst Math Studies Soc Sci, Stanford Univ, June 1974 (*a*)

——— "Homothetic Production Functions" Unpublished memorandum Stanford, Calif Stanford Univ Press, October 1974 (*b*)

———, "Econometrics of Monotonicity, Convexity and Quasiconvexity" *Econometrica*, in press, 1976

Lawrence, Anthony G "Inter-Fuel Substitution Steam Electric Generation's Demand for Fuels" Research Discussion Paper no 8 Washington Res Division, Office Prices and Living Conditions, U S Bur Labor Statis, December 1972

Ling, Suilin *Economies of Scale in the Steam-Electric Power Generating Industry An Analytical Approach* Amsterdam North-Holland, 1964

Lomax, K S "Cost Curves for Electricity Generation." *Economica* 19, no 74 (May 1952) 193–97

McFadden, Daniel "Notes on the Estimation of the Elasticity of Substitution." Working Paper no 57 Berkeley Inst Bus and Econ Res, Univ California, December 1964

Nerlove, Marc. "Returns to Scale in Electricity Supply." In *Measurement in Economics. Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*. Stanford, Calif.: Stanford Univ. Press, 1963.

Nordin, J A "Note on a Light Plant's Cost Curves." *Econometrica* 15, no. 3 (July 1947) 231–35

Oberhofer, W., and Kmenta, J. "A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models." *Econometrica* 42, no 3 (May 1974). 579–90

Spann, Robert M "Rate of Return Regulation and Efficiency in Production: An Empirical Test of the Averch-Johnson Thesis." *Bell J. Econ. and Management Sci.* 5, no 1 (Spring 1974). 38–52

U.S. Bureau of the Census. *Statistical Abstract of the United States. 1973.* Washington. Government Printing Office, 1973.

U.S. Federal Power Commission. *Statistics of Privately Owned Electric Utilities in the United States: 1972* Washington. Government Printing Office, 1973 (*a*)

——— *Steam-electric Plant Air and Water Quality Control Data for the Year Ending December 31, 1970.* Washington: Government Printing Office, 1973 (*b*)

——— *Steam-electric Plant Construction Cost and Annual Production Expenses Twenty-fifth Annual Supplement—1972.* Washington Government Printing Office, 1974.

Zellner, Arnold "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *J Amer Statis Assoc.* 57, no. 298 (June 1962) 348–68.

Money and Economic Activity in the Open Economy: The United Kingdom, 1880–1970

P. D. Jonson

International Monetary Research Project,
London School of Economics, and Reserve Bank of Australia

This paper presents a highly aggregated structural macroeconomic model of the U K. economy The specification is based on the insights provided by recent theoretical work on the working of open economies, and provides the most complete answer yet offered on the reverse causation controversy in monetary economics The key empirical findings include relatively high interest rates and price elasticities in the relevant behavioral equations and a significant disequilibrium real balance effect on a broad expenditure aggregate The results of simulation analysis of the model are reported and throw light on Britain's relatively poor postwar macroeconomic performance

I. Introduction

Postwar British macroeconomic performance has in many ways been dismal Inflation has on average exceeded that of the U K 's major trading partners, and macroeconomic policy reversals have resulted in recurrent policy "stops" as the economy runs up against the balance-of-payments constraint, followed by periods of "go" which sow the seeds of

This paper benefited from helpful comments received when earlier versions were presented at the London School of Economics, Manchester University, the University of Chicago, the Federal Reserve Board in Washington, and at the 1975 Konstanz seminar on Monetary Theory and Policy Particular thanks go to H G Johnson, J M. Parkin, A K Swoboda, B Friedman, H Konig, and C R Wymer, whose advice was especially helpful and whose suite of computer programs provided an essential input I would also like to acknowledge with thanks the benevolence of my employer, the Reserve Bank of Australia, for granting me the opportunity to work at the London School of Economics, and the stimulus provided by members of the International Monetary Research Project, financed by the Social Science Research Council, although the opinions and conclusions are, of course, my own and in particular are not necessarily shared by my employer The assistance of the Science Research Council's Atlas Computer Laboratory, which provided the computing facilities, is also gratefully acknowledged

[*Journal of Political Economy*, 1976, vol 84, no 5]
© 1976 by The University of Chicago All rights reserved

the following stop. A sizable devaluation in 1967 provided the precedent for further devaluations via a floating or, more precisely, sinking exchange rate in recent years. This paper presents a highly aggregative macroeconomic model which explains the trends in British prices, output, and the balance of payments over a relatively long historical period. The model is used to throw light on the reasons for Britain's relatively poor postwar performance, and the analysis suggests ways in which this could be improved.

To state the major conclusions in advance, simulation analysis of the model attributes considerable weight to the relatively rapid postwar growth in the government sector, which has produced persistent pressure in domestic goods markets, raising British prices above purchasing-power parity and producing persistent trade deficits The monetary consequences of the strong growth in the government sector have further added to domestic demand pressures, as excess money balances have increased private sector expenditures and produced the perennial British balance-of-payments deficits These results depend on the findings, reported below, that the responses of the U K private sector, as represented by the price and interest elasticities in the various behavioral equations and the real balance effect on expenditure, are economically appropriate and statistically significant This finding is largely at variance with the conventional wisdom embodied in existing macroeconometric models of the British economy, and so needs emphasis

By way of further introduction, a brief discussion of current knowledge of the role of monetary disturbances is presented This indicates the areas in which the model presented in this paper makes its contribution

Much recent work on modeling macroeconomic relationships has been to include monetary factors, in response both to the worldwide inflationary takeoff and to the challenge provided by the reduced-form econometric exercises presented by Friedman and Meiselman (1963) and many others This work has generated considerable controversy about the econometric methods used, centered on the reverse causation issue Although Friedman and Schwartz (1963) met this objection in advance by arguing on historical grounds for the exogeneity of monetary change in major episodes, they conceded that for minor cycles reverse causation is possibly of equal importance to the influence of money on nominal income, a finding supported by Cagan's (1965) study of the money supply process. This leaves considerable room for debate about the precise role of monetary disturbances in recent experience, especially as none of the existing studies has fully come to grips with the role of endogenous policy reaction.[1] A variety of transmission mechanisms—interest rate effects on investment, wealth effects on consumption, and availability effects based on notions of disequilibrium mechanisms of various types—are incorporated in the big U.S models,[2] and Moroney and Mason (1971) provide a small-scale approximation to the mainstream macroeconomic model The structure of this model is not, of course, as rich or as well specified as that of the larger models, but the work is of particular interest since simultaneous estimation techniques are applied to a model including a hypothesis about the money supply process, which offers a partial answer on the reverse causation issue (it is partial since policy instruments are taken as exogenous).

---

[1] As pointed out by Davis (1969), among others

Considerable recent theoretical work in monetary economics has gone into the open-economy case, suggesting in particular that domestic monetary disturbances are a major reason for fluctuations in the balances of payments of small open economies, and also emphasizing that the law of one price must dominate the behavior of commodity prices (adjusted for changes in the exchange rate) in the long run [3] In pointing out that the domestic money supply of an open economy which maintains fixed exchange rates is adjusted to demand through the balance of payments, this work casts doubt on the findings of Friedman and Meiselman–type tests for open economies In effect, the monetary approach to the balance of payments presents a powerful reason why reverse causation is likely to be particularly important, and suggests that any influence of domestic monetary disturbances on price or income will be strictly a disequilibrium phenomenon The empirical studies generated by this approach[4] have so far concentrated on explaining the balance of payments (or the capital account on the grounds that the current account is largely determined by "real" factors), and use of reduced-form econometric methods means that such studies can at best confirm that "money matters" (for the balance of payments of open economies) Objection to some of the assumptions in these studies, especially to the popular assumption that adjustment is more or less instantaneous, suggests the need for structural models which can answer questions about how and why money matters, as in the closed-economy context The large open-economy models have, however, lagged seriously in their incorporation of money factors,[5] and the standard finding of low or zero price and interest rate elasticities in trade and international capital flow equations leaves unanswered the question of how, if at all, the monetary result obtains in the open economy

The model in this paper offers answers to some of the outstanding questions about the interactions of money and economic activity in the open economy, and pays particular attention to the role of fiscal and monetary policy in the money supply process. It incorporates a Bank rate reaction function, to provide the most complete answer yet given to the reverse causation problem[6] Its major contribution is to develop the dynamics of the monetary model,[7] and a crucial finding is a powerful and significant disequilibrium real balance effect on a broad expenditure aggregate. This finding is especially important since the real balance effect is sufficient to reconcile classical and Keynesian macroeconomics,[8] and conventional and monetary models of the balance of payments.[9] Other important findings include high price and interest rate elasticities and relatively long lags in the adjustment to exogenous shocks, the implications of which will be discussed below.

---

[2] Fisher and Sheppard (1972) provide a survey

[3] Johnson and Frenkel (1975) provide an extensive discussion and much of the recent literature

[4] See, e g , Kouri and Porter (1974), Courchene and Singh (1975), and Genberg (1975)

[5] Ball et al (1975) provide a discussion of the London Business School model of the U.K. economy

The plan of the paper is as follows Section II discusses trends in British prices, output, the balance of payments, and other economic magnitudes. Section III specifies a macroeconomic model based on the insights provided by recent developments in monetary theory, and confronts this model with the data generated by 90 years of U K economic history Section IV discusses the implications of the results for the debate on postwar British economic performance

II. Major Macroeconomic Magnitudes: An Overview

The variable that most closely approximates the aggregate discussed in theories of the balance of payments is described as the "total currency flow" in the official British statistics, and the cumulated value of these flows is designated "reserves" in this study. Changes in reserves are financed by changes in "official reserves" (gold and convertible currencies) and currency swaps, loans from and repayments to the International Monetary Fund, and other exogenous capital flows which do not, proximately at least, influence the money stock Figure 1 illustrates changes in "reserves" and in "official reserves" over the sample period The comparison illustrates a marked contrast in their comparability in

---

[6] The model is estimated with a powerful full-information maximum-likelihood technique developed by C R Wymer which disposes of the simultaneity problem The technique also involves a consistent treatment of the problem of estimating a differential equation system with discrete data, and allows the use of restrictions within and across equations which are suggested by economic theory

[7] Theoretical analyses of various aspects of the dynamics in models which utilize a monetary approach to the open economy have been presented by Parkin (1972, 1974), Mussa (1974), Jonson and Kierzkowski (1975), and Laidler (1975). Knight and Wymer (1975) present an alternative empirical implementation of the monetary approach using postwar U K. data.

[8] Archibald and Lipsey (1958) and Patinkin (1965).

[9] Jonson and Kierzkowski (1975) It is not theoretically necessary, since other channels for the influence of monetary policy on economic activity, such as wealth effects or interest-rate effects, could also be important.

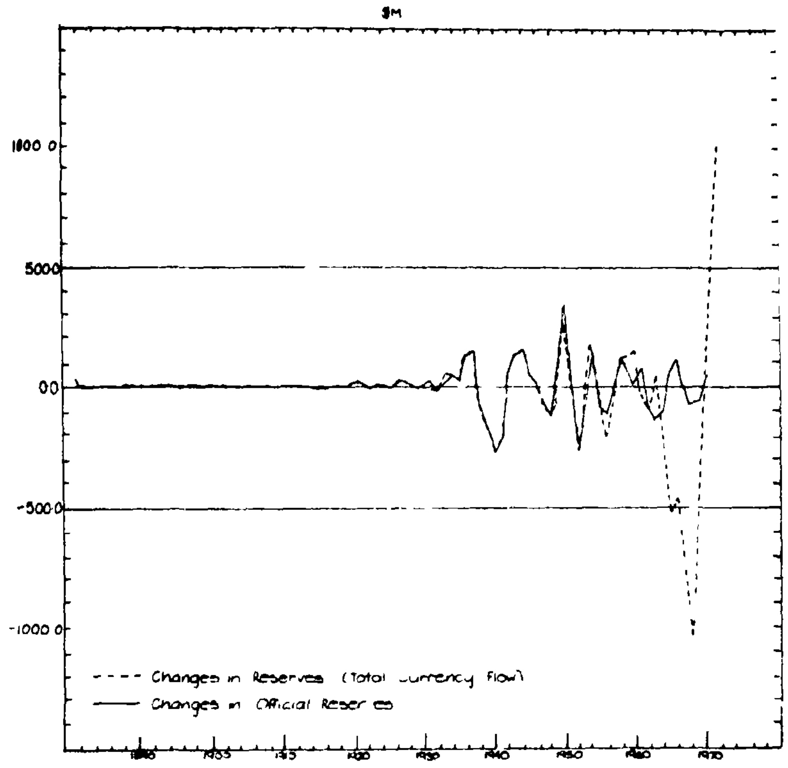


Fig 1 —Changes in reserves and in official reserves

the periods before and after the early 1930s, with fluctuations in both, but especially in reserves, becoming much larger in the second period, particularly in the 1960s Figure 2 illustrates the cumulated value of total currency flows (reserves) and the level of official reserves, which show broadly similar patterns of change except toward the end of the period Growth in reserves was much steadier in the first 50 years of the period than in the second 40, although growth was about a trend that was more rapid in the postwar period than in the earlier years, at least until the mid-1960s.

With regard to the components of the balance of payments, figure 3 illustrates the current balance and net capital flows This shows that the historical tendency for the United Kingdom to export capital was reversed in the 1930s, during World War II, and in a few years of the postwar period. The obverse of the historical tendency, at least in the earlier years when reserves did not fluctuate wildly, was a tendency toward current account surpluses, rising to a peak in the early years of

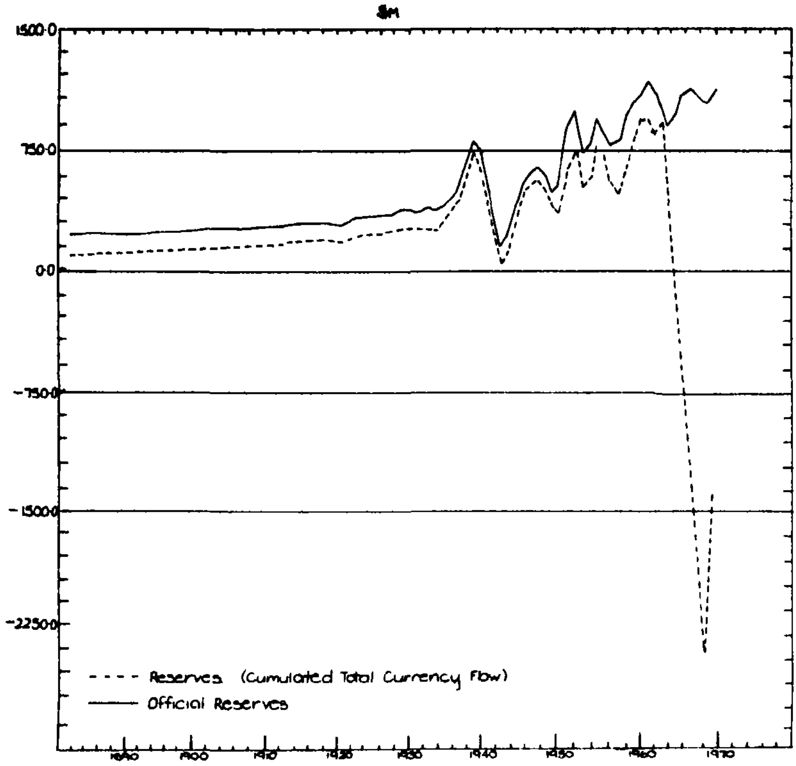


FIG. 2.—The levels of reserves and official reserves

the twentieth century and again in the early 1920s. In the 1930s and World War II, large current account deficits mirrored sizable net capital inflows. In the postwar years, however, the current and capital accounts have often moved in the same direction, in marked contrast to the earlier pattern.

Figure 4 illustrates the volume of exports and of imports of goods and services. A relatively steady growth in exports before World War I and after World War II was interrupted by both wars and in the interwar period. The average annual rate of growth in the volume of exports from 1952 to 1970 was 4.2 percent, somewhat higher than the 3.4 percent in the period 1883–1913; in addition, fluctuations were smaller in the postwar period than in the earlier years. Fluctuations in the value of exports tend to be greater than those in volume, reflecting a tendency for parallel movements in price and quantity. The quantity of imports has tended to grow more steadily than exports, although import growth was also slowed by the two wars. The average growth rate of imports is

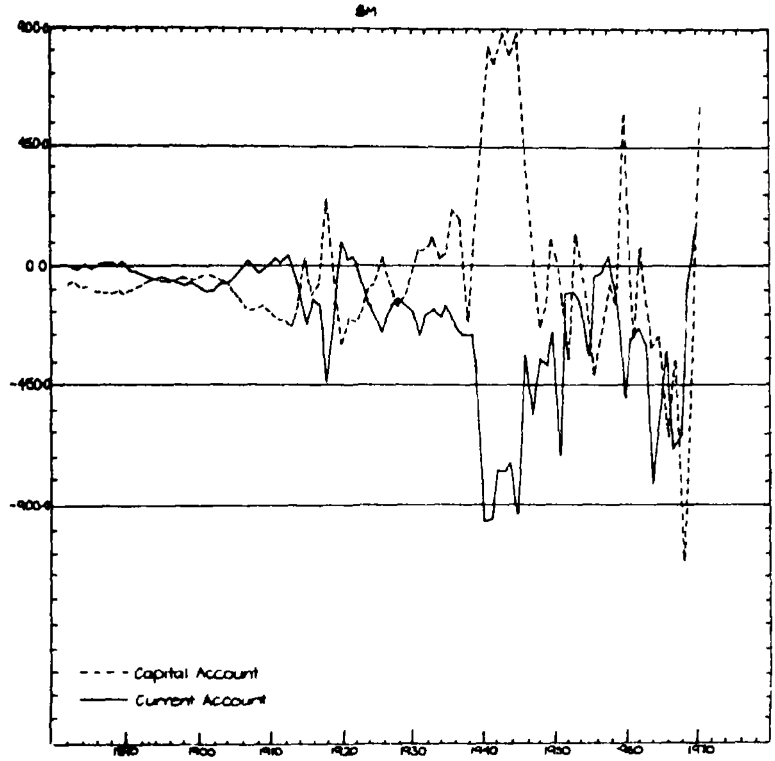


Fig 3 —Net capital flows and the current balance

considerably higher than that of exports, a major arithmetical cause of the change from a strong balance of payments in the early years of the sample to the tendency toward deficits in the later years Strong price fluctuations tend to be inversely correlated with fluctuations in import volume. The other components of the current balance are not illustrated, net property income from abroad grows fairly steadily, with checks caused by liquidation of overseas assets during the two wars, especially World War II, net grants and transfers are small and relatively unimportant in their impact on the balance of payments

Output, expenditure, and current government spending are illustrated in figure 5. Aggregate output grew at a relatively steady rate before World War I and even more steadily, and at a slightly greater rate, after World War II. Output was boosted by the war efforts (reflected strongly in the government-spending series) and fell considerably after both wars. The major decline in output occurred in the depression of the 1930s, although there were also significant reductions of output in 1882–83,

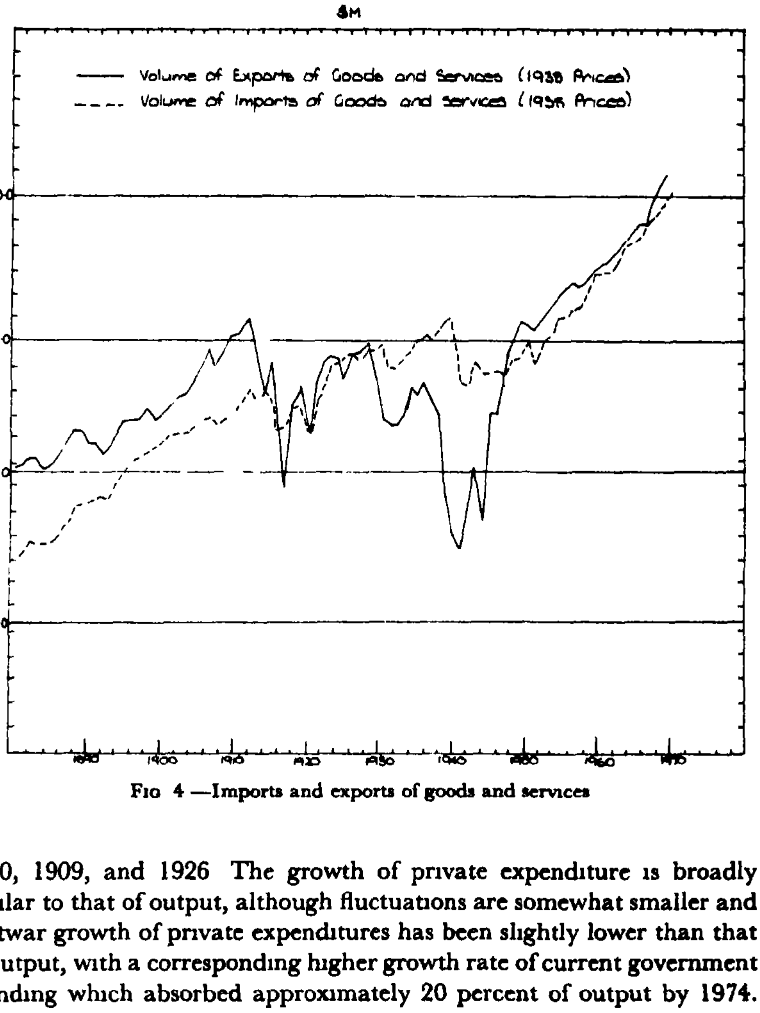


FIG 4.—Imports and exports of goods and services

1900, 1909, and 1926. The growth of private expenditure is broadly similar to that of output, although fluctuations are somewhat smaller and postwar growth of private expenditures has been slightly lower than that of output, with a corresponding higher growth rate of current government spending which absorbed approximately 20 percent of output by 1974. (Current government spending absorbed only 5 percent of output in 1880, and this proportion had risen to 8 percent by 1913 and to 11 percent by 1938.) The effect of fluctuations in government spending on the budget deficit, and of the deficit on domestic credit expansion, is clearly visible in figure 6.

Figure 7 illustrates U.K. and U.S. prices—the latter representing world prices in this study[10]—and shows a remarkable degree of conformity once

---

[10] U.S. prices, income, and short-term interest rates are assumed to represent the corresponding world magnitudes in this study. This is clearly an approximation, although Fisher's (1912) discussion of the dominant position of the U.S. economy at the turn of the century provides some justification for the assumption.

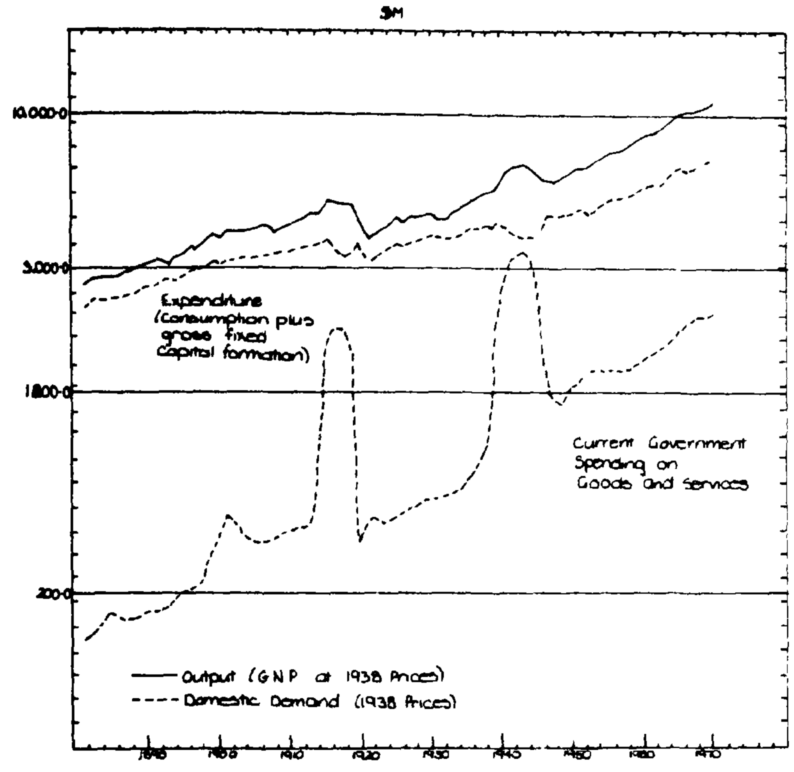


FIG. 5 —Output and domestic demand categories

changes in the exchange rate are considered The major divergences from purchasing-power parity occur in the 1890s and the 1930s, when British prices did not fall so rapidly as U S prices

Figure 8 shows the growth of money and bonds over the sample period Money grew far more steadily than bonds, with rapid increases during both wars and the only major fall in nominal balances occurring in the 1920s. Bonds were redeemed on average during the 1890s and in the first decade of the twentieth century, when budget surpluses were the rule, and rose extremely rapidly during both wars Some additional light is thrown on these movements in figure 9, which shows that short-term U S. interest rates were well above U K rates until the 1930s (although the gap narrowed considerably during World War I) and fell more than U.K. rates during the cheap money period of World War II United Kingdom interest rates have been on average higher than U.S. rates in the postwar years, and the postwar rise in interest rates is correlated with

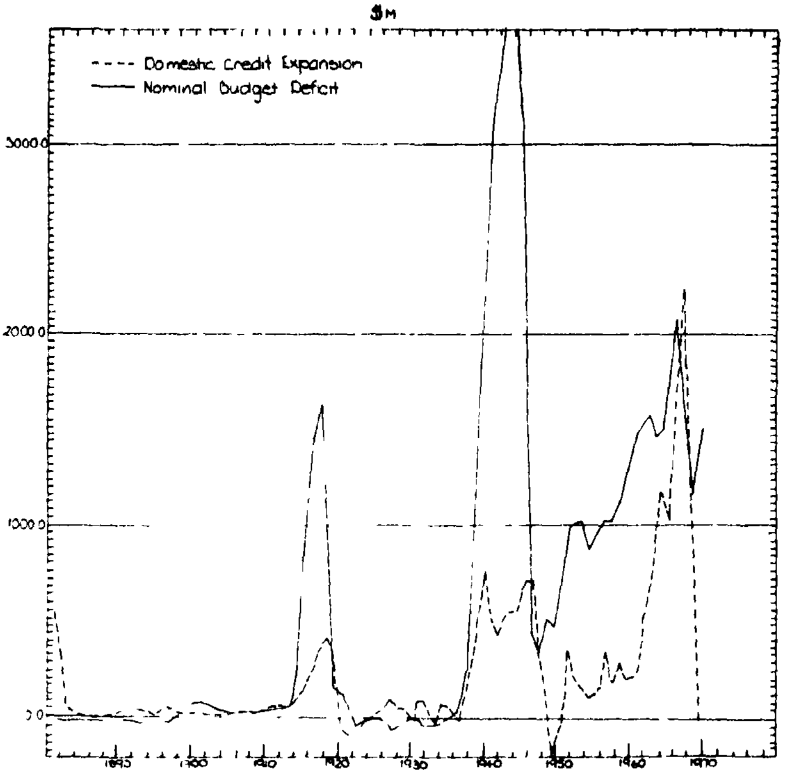


Fig. 6 —Domestic credit expansion and the budget deficit

the rise in the velocity of circulation of money (as are earlier fluctuations in interest rates and velocity)

In conclusion, it is worth observing that the striking empirical regularities apparent even in the raw data illustrate the value of employing a relatively long sample period to test economic theory. The model-building exercise utilizing these data is designed to yield the precise estimates of the elasticities of the various behavioral functions, and of speeds of adjustment, which are a necessary part of scientific knowledge of economic systems.

III. A Monetary Model

In specifying the equations of the model, it is convenient to proceed *as if* each major aggregate is determined, proximately at least, by one group of economic agents: households consume and invest real goods and hold

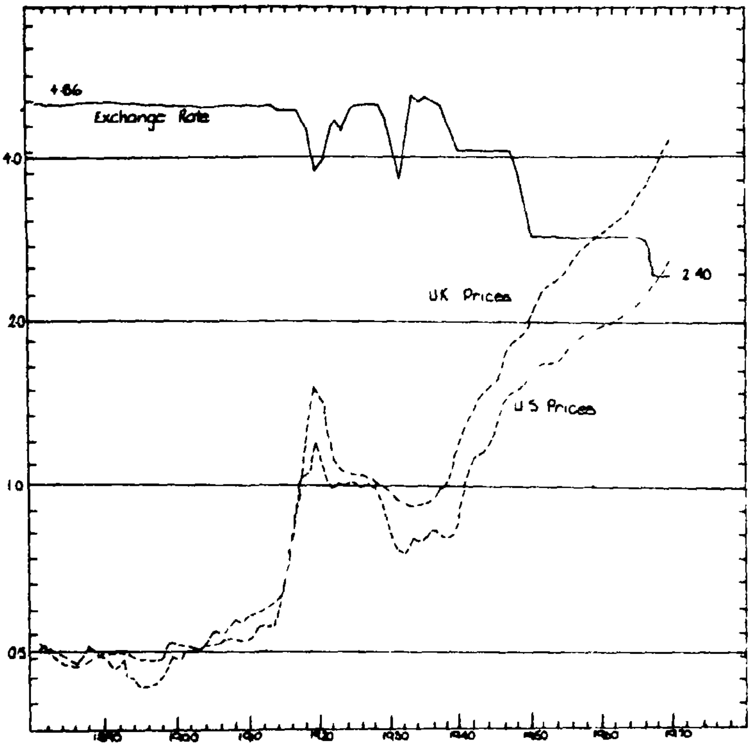


Fig 7 —U K and U S prices and the exchange rate

financial assets, including money, sterling-denominated bonds, and overseas assets, either directly or via their ownership of domestic firms; entrepreneurs determine the volume of output and the price at which it is offered for sale, and the domestic government (including the banks) consumes real goods and services, finances its consumption by the issue of money and bonds and by levying taxes, and sets nominal interest rates (represented by Bank rate) in the short run In addition, foreigners purchase British exports and hold U K assets, although the latter are deducted from the overseas assets of U K. residents and are not treated explicitly in the model

The markets explicitly represented in the model consist of markets for money, bonds, and goods. Conceptually there are three goods: imports whose price is given on world markets, exports, and other domestically produced goods Given the assumption that nominal interest rates and exchange rates are pegged in the short run, the supplies of money and of

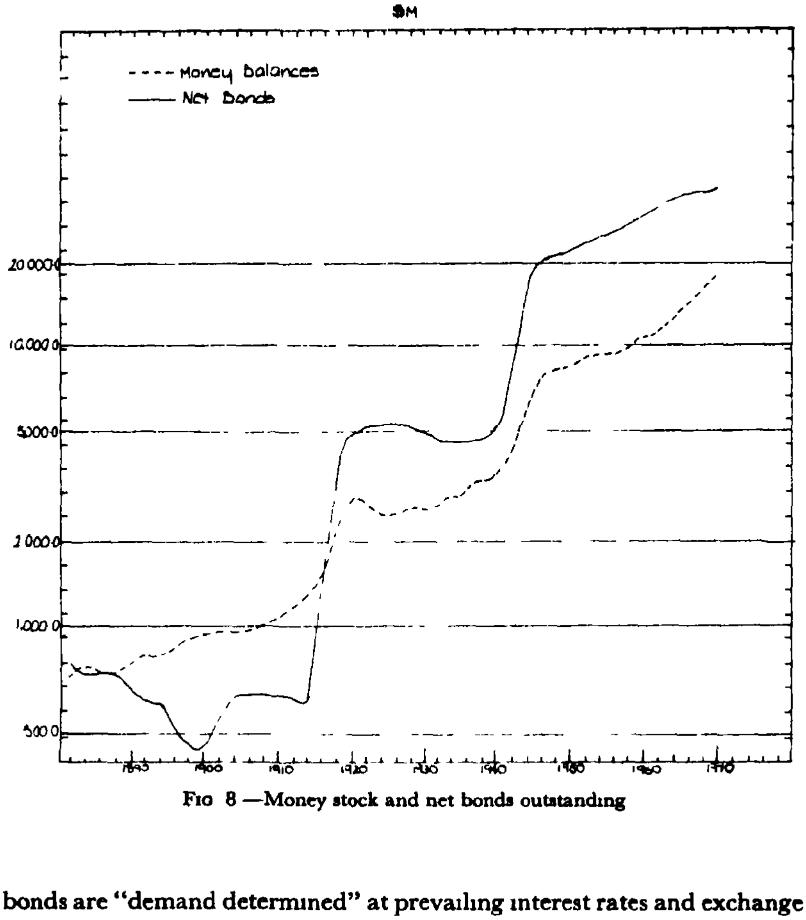


FIG 8 —Money stock and net bonds outstanding

bonds are "demand determined" at prevailing interest rates and exchange rates, which will of course be changed if the government's perceived macroeconomic objectives are not being achieved. The change in the value of net overseas assets (i e , net capital outflows) is given, once the balance of payments is determined, and given the values of exports and imports.

The following subsections discuss more fully the relationships in the model, and then the properties of the model as a whole are discussed.

Aggregate Commodity Demand

The derivation of effective demands for flows of commodities and stocks of assets is in two steps: the determination of full equilibrium or long-run demands, and the adjustment processes linking actual and equilibrium demands. The full equilibrium demand for domestic expenditure *d* is

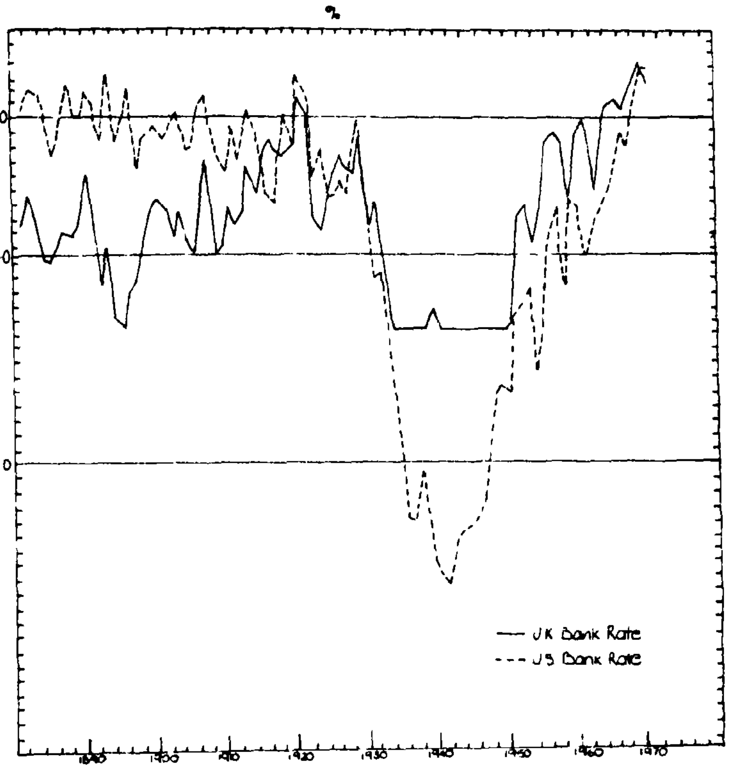


Fig 9 —U K and U S interest rates

assumed to be a function of disposable income ($y - T/p$) and Bank rate (r_b), the latter representing any real interest rate effect on expenditure,[11] on the assumption that the expected rate of change of prices was zero, or constant, for most of the sample period [12]

$$d = d_0(y - T/p)^{\beta_1}(r_b)^{\beta_2} \tag{1a}$$

---

[11] Dale Henderson has correctly pointed out that the current specification is not fully consistent with the standard stock adjustment model of investment in consumer durables and industrial capital. Since no meaningful alternative can readily be substituted, it is justified in terms of its extreme simplicity and the strategy of concentrating attention on issues that are less well understood than investment and consumption behavior

[12] The normal or long-run expectation for most of British history has almost certainly been of a constant price level, perhaps modified by blips associated with wars Even in recent experience, Carlson and Parkin (1975) find that postwar British price expectations are virtually constant until quite late in the inflationary takeoff of the 1960s, which is consistent with the arguments advanced by Sargent (1971).

It is further assumed that various constraints imposed by the costs of obtaining information, the existence of uncertainty, and the problem of maximizing the time path of consumption in the face of various exogenous disturbances lead to partial adjustment toward the long-run equilibrium position, with the change in expenditure a function of the gap between actual and desired wealth ($\hat{w}$) at any point in time This gives

$$D \log d = \alpha(\log \hat{d} - \log d) + \gamma(\log \hat{w} - \log W/p) \qquad (1b)$$

The asset adjustment term is included to represent the influences of hoarding and dishoarding of financial assets on current expenditure, and arises from a generalization of the Archibald and Lipsey (1958) real balance effect[13] Most theoretical generalizations result in a vector of real asset effects, each with a different coefficient γ[14] In the current model, decisions about real assets are subsumed in the expenditure decision, and only monetary disequilibrium has a significant influence on expenditure when a more general model which also includes the difference between actual and desired bonds is estimated This may be a function of the relatively unreliable nature of the bonds series (since it is derived residually and is the cumulated value of changes in bonds with no explicit allowance for changes in market value), although if it is true that the composition of assets adjusts relatively quickly, monetary disequilibrium (for example) may be a good proxy for financial asset disequilibrium in general This relatively neglected issue is discussed further below

In estimating the model, equation (1a) and the determinants of $\hat{w}$ are substituted into equation (1b). A wartime dummy variable is also added to equation (1a)[15] to correct for any influence of the wartime controls and rationing on expenditure behavior

The volume of exports is assumed to be determined in the long run by world demand at prices given by U K producers The long-run demand for exports is assumed to be a simple function of world income (y_w), represented by U.S. real income, and the price of exports (P_x) relative to that of U K imports (P_l), a proxy for world traded-goods prices in sterling. Actual (rather than disposable) world income is used because export demand includes foreign government demand.

$$\hat{x} = x_0 y_w^{\beta_1} (P_x/P_l)^{\beta_2}. \qquad (2a)$$

---

[13] An equation very similar to (1b) is derived by Dornbusch and Mussa (1975) from a model in which consumption is optimized over time in a world in which money is the only asset

[14] The model developed by Parkin et al. (1973) in effect generalizes in terms of a set of different γ's (although consumption is determined residually in their model).

[15] And elsewhere where appropriate. It should be noted that wartime dummies are the only "special factors" allowed for in the model.

Exports are also assumed to adjust so as to remove the discrepancy between actual and equilibrium exports at a constant proportional rate, with the speed of adjustment influenced by a short-run supply constraint variable, defined as the ratio of domestic output (y) to its steady-state growth path ($y^* e^{\lambda_1 t}$) This provides a further test of the popular notion that actual exports are reduced when output is growing slowly, and vice versa.[16]

$$D \log x = \alpha(\log \hat{x} - \log x) + \beta \log (y/y^* e^{\lambda_1 t}) \tag{2b}$$

Real government spending is assumed to be exogenous in view of the changes in the theory and practice of fiscal policy which have taken place over the past century and the great influence of wartime expenditure on the series

Aggregate Commodity Supply and Price Determination

On the supply side, the usual Keynesian assumption that output equals aggregate demand less imports (after allowing for stock building) is modified. Expected sales (S^e) are divided between domestic production (y) and imports (i) in proportions that depend on the relative price of imports and domestic output (P_i/P), and actual imports and production are assumed to adjust to expected sales at constant although different rates[17] A partial response to expected sales is justified by uncertainty about future demand and the operation of short-run capacity constraints Symbolically, the import-output decision is represented by

$$\hat{i} = i_0(P_i/P)^{\beta_1} S^e, \tag{3a}$$

$$D \log i = \alpha_1(\log \hat{i} - \log i), \tag{3b}$$

$$\hat{y} = [1 - i_0(P_i/P)^{\beta_1}]S^e, \tag{4a}$$

$$D \log y = \alpha_2(\log \hat{y} - \log y) \tag{4b}$$

Expected sales are assumed to equal current demand, modified by a dummy variable ($Q1$) to allow crudely for the depression widely expected at the end of World War I and the separation of Ireland

$$S^e = (d + x + g)e^{\beta_2 Q1}$$

Although this specification might be said to be still "demand dominated" to an unreasonable degree, it should be noted that the steady-state rate of growth of expected sales will equal that of resources

---

[16] Ball, Eaton, and Steuer (1966) previously tested for this effect of U.K. exports Norton (1969) applies the same idea to U K financial markets

[17] Bergstrom and Wymer (1974) develop a version of this model which also takes explicit account of stock-building behavior, although the price elasticity is constrained to be -1 Anderson, Pearce, and Trivedi (1975) use a similar model (ignoring price effects) to explain output decisions with much more disaggregated data

Empirically, the above specification is preferable to one[18] in which sales are expected to grow at a constant rate equal to the growth rate of resources, although there could well be important returns to a more sophisticated analysis of expected sales.[19]

The proposition that domestic prices adjust toward purchasing-power parity is an integral part of the monetary approach, although the possibility of running down or building up reserves, changing trade and capital controls or taxes and other short-run market imperfections are likely to produce delays in the adjustment of domestic to world prices, the role of nontraded goods having received considerable attention in the theoretical literature The widely tested expectations-augmented Phillips curve (applied to prices rather than to wages) provides the basis of an equation for the price of output in the current model which is consistent with the monetary result in the long term but allows deviations from purchasing-power parity during the adjustment period Specifically, if it is assumed that world prices adjusted for the exchange rate (EP_w) are the prices rationally expected to prevail in the long run, that the price of output adjusts toward this with a lag which incorporates various real world frictions, and that some measure of the pressure of demand relative to full employment supply ($z/s^*e^{\lambda_1 t}$) influences the rate of inflation during the adjustment period, the following equation is obtained

$$D \log P = \alpha[\log EP_w - \log P + \beta \log (z/s^*e^{\lambda_1 t})] \tag{5}$$

When this equation is estimated with $z = d + x + g$, it yields, as expected, a positive and significant estimate of β. Better results are obtained, however, if the different components of demand are allowed a separate influence on price adjustment. In particular, fluctuations in government spending (g) and in other domestic expenditure (d) appear to exercise a much stronger influence than changes in exports (x), which in fact have an approximately zero direct influence on price adjustment[20] This is broadly consistent with the fact that domestic producers have a much stronger short-run monopoly position in domestic markets than in world markets, and suggests an interesting way in which to generalize the Phelpsian[21] literature for the open-economy case

Two alternative export price equations were tested: one analogous to that for the price of output in general, and another, somewhat simpler equation in which export prices adjust slowly to the trend of traded-goods prices and immediately to changes in the prices of traded goods and of domestic output (which are the prices in closely related markets) The second specification, which is effectively the same as the first (since demand effects are included indirectly via their effects on changes in domestic prices), performs somewhat better, and so is included in the model:

---

[18] Suggested by A L. Marty.

[19] Lucas (1973) investigates the role of divergences between actual and expected prices and the influence of the variance of government policy on the price-output decision A simple attempt to incorporate the first of these effects was only of marginal success and is not reported here

[20] There are indirect effects via the income effect on domestic expenditure.

[21] Phelps et al. (1970) explore the notion of short-run monopoly positions which are eroded over time to provide microeconomic underpinnings for wage and price equations.

$$D \log P_x = \alpha(\log P_t - \log P_x) + \beta_1 D \log P_t + \beta_2 D \log P \qquad (6)$$

Demand for Financial Assets

The private sector's holdings of financial assets are disaggregated into money (M), sterling-denominated "bonds" (B), and foreign assets (S) The model includes behavioral equations for the first two, and the third is determined residually "Bonds" is in fact a composite asset, consisting mainly of the nonmonetary liabilities of the government sector held by the public, less bank advances to the private sector, although it is treated for empirical purposes as if it were a homogeneous internationally traded asset

The desired stocks of money and of bonds are assumed to be functions of wealth or income (y) and the return on domestic and foreign bonds, represented by Bank rate (r_b) and U S Bank rate (r_w), adjusted for the expected change in the exchange rate ($\dot{E}^e$)

$$\hat{m} = m_0 y^{\beta_1}(r_b)^{\beta_2}(r_w + \dot{E}^e)^{\beta_2}, \qquad (7a)$$

$$\hat{b} = b_0 y^{\beta_3}(r_b)^{\beta_4}(r_w + \dot{E}^e)^{-\beta_4} \qquad (8a)$$

The interest elasticities are restricted to the equal or equal and opposite (except for an additive constant) in line with the view that domestic and foreign bonds are perfect substitutes in the long run [22] Various simple proxies for the expected rate of change of prices were also included in these equations (and in eq [1]), to capture any influence of price changes not represented in nominal interest rates or in $\dot{E}^e$, but such variables were always insignificant Negative results are also obtained when estimates of net financial wealth are used as an alternative to or in addition to permanent income[23] as the constraint variable, although this result is not especially strong, as the net wealth series is extremely crude.

The dynamics are handled very simply, with a first-order partial adjustment model in which the change in the nominal value of money and

[22] The data appear to satisfy this restriction in the sense that, when the model is estimated without it, the separate coefficients are not significantly different, although the standard errors are quite large

[23] Proxied in the usual way by a distributed lag on actual income

bonds is proportional to the discrepancy between desired and actual stocks of real assets:

$$D \log M = \alpha_1(\log \hat{m} - \log M/p), \tag{7b}$$

$$D \log B = \alpha_2(\log \hat{b} - \log B/p) \tag{8b}$$

In interpreting the implied lags ($1/\alpha$ provides an estimate of the mean time lag), one should recall that it represents the average effect of lags in the formation of expectation about income or interest rates, the cost of adjustment, and optimal savings behavior The result obtained for the full period to 1970 is that α_1 is considerably larger than α_2 This implies that the *composition* of portfolios adjusts slowly, a finding that is puzzling to some economists[24] Various alternative specifications yield the same conclusion for the full sample period, although for the period to 1939, α_1 is approximately equal to α_2, an interesting finding which is discussed further when the parameter estimates are presented.

It is perhaps worth stressing that the model is constructed so that money and bonds (and other assets) grow in fixed proportions in the steady state, and that any anticipated trend rate of growth of wealth or permanent income is effectively built into the constant term in partial adjustment models such as those utilized in this study[25]

The Money Supply Process and the Balance of Payments

As noted above, U K. monetary policy is conducted by setting some important nominal interest rates Part of any consistent description of the workings of the U.K economy must therefore include a policy reaction function[26] to explain changes in nominal interest rates, represented by Bank rate (r_b) in the current model. In addition to the offensive policy reactions represented by the deviations of output, prices, and the balance

---

[24] See, e g , Frenkel and Rodriguez (1975), who distinguish carefully between overall asset decumulation, which is assumed to take time, and more or less instantaneous portfolio reallocations in a theoretical "portfolio" model of the balance of payments.

[25] Substituting eq. (7a) into (7b), ignoring for simplicity the interest rate terms, yields

$$D \log M = \alpha(\log m_0 + \beta_1 \log y - \log M/p).$$

This is observationally equivalent to a model in which the proportionate rate of growth of money balances equals a trend term λ plus an adjustment term indicating the response to current disequilibrium.

$$D \log M = \lambda + \alpha(\log m_0' + \beta_1 \log y - \log M/p),$$

where $m_0' = m_0 e^{-\lambda/\alpha}$. A similar comment applies to most of the other equations in the model and in other, similar models (Bergstrom and Wymer 1974)

[26] Jonson (1974) provides a survey of the growing literature on this subject and an analysis of Australian stabilization policy in the 1960s. As noted in the introduction of the current paper, one of the major unresolved issues in the debate on reverse causation is the effect of incorporating policy reactions in models measuring the interactions of money and economic activity.

of payments from target levels, Bank rate is also assumed to react defensively to changes in U S. short-term interest rates·

$$
\begin{aligned}
D \log r_b = {} & \alpha \log (r_0/r_b) + \beta_1 D \log r_w + \beta_2 \log (y/y^* e^{\lambda_1 t}) \\
& + \beta_3 \log (P/EP_w) + \beta_4 \log (R/R^* e^{(\lambda_1 + \lambda_2)t}) \\
& + \beta_5 \log (L/L^* e^{(\lambda_1 + \lambda_2)t}).
\end{aligned} \tag{9}
$$

The target levels for the major macroeconomic variables are assumed to be consistent with the steady state of the model. Thus, for example, deviations of output (y) from its steady-state growth path defined by $y^* e^{\lambda_1 t}$, rather than the level of output, represent the "internal balance" objective, similarly, purchasing-power parity (EP_w) represents a rational price level objective

In contrast to Bank rate policy, fiscal policy is assumed to be exogenous, as noted above in conjunction with government spending Thus, tax receipts are related simply to the value of aggregate output, with given tax rates (a trend term allows for the progressive nature of tax schedules):

$$T = T_0 e^{\beta_1 t} Py \tag{10}$$

Domestic credit expansion (DC) is determined by the budget deficit ($Pg - T$) less the net change in the value of "bonds" held by the private sector.[27]

$$DC = Pg - T - DB \tag{11}$$

This decomposition arises from a model in which the government and banking sectors are aggregated In common with the more complex models developed recently, particularly by Brunner and Meltzer (e.g., 1974), it allows for a major and largely exogenous impact of fiscal policy and for the effect of changes in domestic and foreign interest rates In contrast to this work, however, the present model abstracts from the role of the private banking system, employing a "portfolio" approach rather than a "monetary multiplier" approach to the money supply process.

The model is completed by an identity determining the change in reserves as the difference between the flow demand and supplies of money·[28]

$$DR = DM - DC \tag{12}$$

It should be noted that the model determines the sterling value of changes in reserves, but not the allocation of a change in their sterling value between an exchange rate change and a change in the quantity of reserves. In principle, it would be possible to model official intervention in the foreign exchange market and determine what proportion of a given change in the sterling value of reserves is taken in the dollar value of reserves (Q) and what proportion in a change in the exchange rate (E), since $DR = D(EQ) = DE \cdot Q + E \cdot DQ$. In practice, the relevant coefficients would be difficult if not impossible to determine (and in any case, the relationship might be strongly influenced by largely exogenous factors) The estimated model could in principle be used to determine the behavior of a freely floating exchange rate by assuming $Q = \bar{Q}$ and solving for E

---

[27] Recall that "bonds" is a composite asset consisting mainly of nonmonetary government debt held by the private sector, less the value of bank advances to the private sector

[28] In the estimated model, this identity contains the change in another term, "loans" (L), which includes official revaluation gains and losses and the value of loans from and subscriptions to the IMF, and currency swaps with other central banks, assumed to be exogenous.

The model as a whole is summarized in Appendix A, and the variables and other notation, along with their definitions, are given in Appendix B. It has a steady-state solution which is relatively simple when $\beta_2 = 1$ and $\beta_{24} = 0$·

$$\begin{aligned} X &= X^* e^{\lambda_1 t}, \qquad & X &= d, x, i, y, \\ Y &= Y^* e^{\lambda_2 t}, \qquad & Y &= P, P_x, \\ Z &= Z^* e^{(\lambda_1 + \lambda_2)t}, \qquad & Z &= T, M, B, C, R, \\ r_b &= r^* \end{aligned}$$

That is, in the steady state the real values of expenditure, exports, imports, and output all grow at the same rate, determined by the growth rate of world income and real domestic resources[29] (λ_1) Domestic prices grow at the rate (λ_2) at which world prices grow, and the nominal variables in the system, taxes, money, bonds, domestic credit, and official reserves grow at a rate ($\lambda_1 + \lambda_2$) equal to the rate of growth of nominal income This steady-state solution is derived on the assumption that all the exogenous variables grow at the appropriate constant rate and that certain parameters, notably the income elasticities of the demands for expenditure, exports, money, and bonds, and the nominal income elasticity of the tax function, be equal to unity This provides a cross-equation restriction which increases the efficiency of the estimator

The model is simple enough so that its properties can be discussed to a considerable extent independently of the parameter estimates and compared with those of alternative models The following heuristic account of the model's response to changes in the exogenous variables illustrates its properties Figure 10, which presents a simple schematic representation of the structure of the model, may be helpful in tracing the interrelationships discussed in the following subsections. Note that the dotted line represents a linkage—from changes in U.K. holdings of U.S. dollars to world prices via any effect on the world money stock—not

---

[29] Real government spending must also grow at the rate λ_1

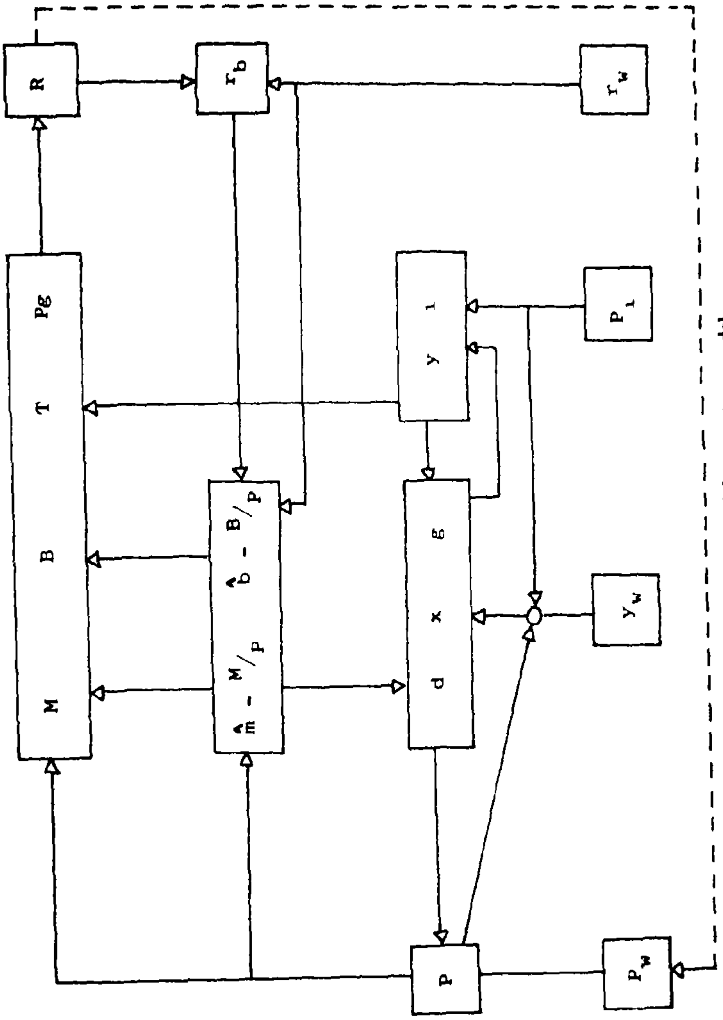


FIG 10 —Structure of the monetary model

included in the present model, and that not all the linkages in the model are represented in this diagram (e g , the effect of a change in taxes on disposable income and expenditure is excluded to avoid cluttering the diagram)

A Rise in World Prices (including Import Prices)

This sets two offsetting forces in motion exports and domestic output rise and imports fall (relative to what they would have been without the change) as the change in relative prices induces a shift in world demand in favor of U.K. goods and as U K. suppliers find it cheaper at the margin to produce rather than import; on the other hand, U.K. prices begin to rise, directly through the competitive or expectational variables in the price equations and indirectly through the increased pressure on domestic resources The dynamic response is complicated by the fact that the rise in domestic output adds directly to domestic demand, although the rise in income also induces a rise in desired money balances which tends to induce offsetting reductions in domestic expenditure, and this is reinforced by the tendency for the rise in prices to reduce the real value of existing money balances. The rise in income and prices and the associated buildup of nominal money balances and bonds tend to strengthen the balance of payments, although this effect is offset by the increased nominal government spending on the assumption that real spending is unchanged.[30] Since tax receipts are also rising along with nominal income, however, domestic credit expansion is likely to proceed less rapidly, and the net short-run effect on the balance of payments is almost certainly favorable. This parallels the predictions of both the simpler monetarist model in which the balance of payments unambiguously improves because of the accumulation of real balances induced by the price rise, and of simple conventional models which concentrate on the initial improvement in the current account

An Increase in U K. Government Spending

This leads to an initial increase in the budget deficit which worsens the balance of payments in the short run The effect is not one for one, as in simpler models, since the increase in aggregate demand raises income and thus expenditure, which together with the increased government spending places pressure on domestic resources, with price effects measured by β_9 and β_{10} This has secondary effects on aggregate demand, via the price effects on the demand for exports and via real balance effects, which, by inducing a buildup of financial assets, produces an offsetting improvement in the balance of payments, reinforced by the tendency for tax receipts to rise The offsetting effects take time to operate, however, and if government spending is not subsequently reduced, involving a probably painful period of falling income (or slower growth), the U K economy will find itself with a price level above purchasing-power parity and a persistent deficit on external account Resolution of this dilemma by a change in the exchange rate, analytically equivalent to a rise in world prices, will work provided expectations of further devaluations do not set in, reducing the demand for British financial assets.

---

[30] If real spending expands, domestic credit expansion will increase by more in the short run

An Increase in Bank Rate

Although Bank rate is an endogenous variable in the model, it is legitimate to consider the effect of an exogenous increase which can be interpreted as resulting from a change in a parameter (most simply the constant) in the policy reaction function, or from a suspension of the normal mechanism (the wartime cheap money policy being the prime example in the sample period)[31] The longer-run effect on the balance of payments depends on the interest elasticity of the expenditure function, although in the short run this is negligible relative to the effect on financial assets The impact effect on the balance of payments thus depends mainly on the relative interest elasticities of the demand for money and for bonds and on the relative speeds of adjustment of money and bonds, a rise in Bank rate tends to improve the balance of payments in the short run, since $|\alpha_8\beta_{17}| > |\alpha_7\beta_{15}|$ This is very much in line with traditional accounts of the effects of changes in Bank rate

In summary, the model exhibits all of the features of the standard monetarist model as properties of its steady state, although it has a much richer dynamic structure than can be considered in purely theoretical analyses Indeed, its short-run dynamics in response to various exogenous changes accord more closely with conventional views about the adjustment mechanism than do most existing monetary models

Estimates of the structural parameters are obtained from annual data for the period 1885–1970 using a full-information maximum-likelihood technique for estimating the coefficients of a discrete approximation to a system of differential equations developed by Wymer[32] In estimating the model, the income elasticities are constrained to be equal, but not unity, since it is desirable to test whether or not $\beta_2 = 1$ In addition, β_{24} is not constrained to be zero, since there has been a rising ratio of government spending to total income over the sample period, financed in part by rises in average tax rates The parameter estimates, together with the *t*-ratios derived from the estimates and their asymptotic standard errors, are presented in table 1.

The estimates are generally very plausible, with all signs in accord with prior beliefs and the income elasticities of demand (β_2) not significantly different from unity, although the point estimate is slightly greater Examining the adjustment speeds measured by the parameters α_3, α_4, α_7, and α_8, which can be interpreted as the reciprocal of the mean of an exponentially distributed lag with which the logarithm of each relevant variable adjusts to changes in the variables influencing it, we have

---

[31] This is modeled by assuming that the desired level was a constant amount below what it otherwise would have been in 1940–51 The amount is measured by β_{19} in the model

[32] Wymer (1972) provides a general theoretical discussion of the econometric issues.

TABLE 1

Estimates of Structural Parameters of a Monetary Model, 1885–1970

Parameter	Estimate	t-Ratio	Parameter	Estimate	t-Ratio
α_1	0.2108	3.61	β_{11}	−0.0283	1.51
α_2	0.2880	5.04	β_{12}	0.5199	15.25
α_3	0.4601	5.19	β_{13}	0.8463	9.23
α_4	0.2818	5.04	β_{14}	0.3174	5.31
α_5	0.1341	4.10	β_{15}	−0.2093	12.31
α_6	0.0784	2.48	β_{16}	5.8169	4.24
α_7	0.1874	8.77	β_{17}	2.0102	4.78
α_8	0.0691	4.25	β_{18}	0.3465	3.99
α_9	0.6881	5.64	β_{19}	−0.6199	5.03
γ	−0.1332	4.55	β_{20}	0.3147	1.07
β_1	−0.2393	2.84	β_{21}	−0.5881	3.44
β_2	1.0246	76.09	β_{22}	−0.2331	2.33
β_3	−0.2887	4.44	β_{23}	0.0588	1.21
β_4	0.2298	1.19	β_{24}	0.0100*	
β_5	−0.9325	3.42			
β_6	−3.4216	4.97	λ_1	0.0172*	
β_7	−0.3737	5.11	λ_2	0.0370	2.44
β_8	−0.2731	4.44			
β_9	0.9412	3.52			
β_{10}	0.7525	3.58			

* Value imposed

approximate mean time lags of just over 2 years for imports, 3½ years for output, 5½ years for money balances, and 15 years for bonds.[33] The other adjustment parameters α_1, α_2, α_5, α_6, and α_9 need to be considered in conjunction with the other influences on the relevant variables, thus, exports adjust with a mean lag of 3½ years if output is growing at its steady-state rate, and more quickly if output is growing faster. Expenditure adjusts with an average lag of just under 5 years if the money market is in equilibrium, but expenditure rises more or less rapidly depending on whether or not the private sector has excess money balances, and the estimate of γ indicates that the disequilibrium real balance effect is powerful and significant. The price of output and of exports adjusts on average quite slowly to changes in world prices or the exchange rate (with $\alpha_5 = .13$ and $\alpha_6 = .08$), but since prices also adjust to contemporaneous changes in government spending and domestic expenditure, and to changes in import prices in the latter case, price adjustment is more rapid than the partial equilibrium speeds suggest. The partial equilibrium lag in the adjustment of Bank rate is about a year and a half, although Bank rate also reacts strongly to contemporaneous changes in U S interest rates, as indicated by the estimate of β_{18}, and the targets of stabilization policy.

---

[33] For the period to 1939, $\alpha_7 = .0743$ and $\alpha_8 = .0896$, suggesting that for the period to 1939 the composition of portfolios adjusted relatively quickly, providing empirical support for the conjecture advanced by Frenkel and Rodriguez (1975). As noted above, this conjecture is not supported by the estimates for the period to 1970, and it is possible that, while asset adjustment has tended to become more rapid in the postwar period, as implied by the comparison of the two estimates of α_7, this has been approximately offset in the case of bonds by the extensive postwar controls on the British capital account.

The general impression given by these results is that the economy adjusts rather more slowly than the emphasis in many theoretical contributions suggests In interpreting this result, one should recall that one reason for lags in adjustment when actual values of the explanatory variables are incorporated into the supply and demand functions is the lag in the formation of expectations about the "permanent" variables which are assumed to influence economic behavior, whereas theoretical models often assume that expectations are given If, for example, the 5½-year lag in the adjustment of money balances is interpreted as mainly reflecting a lag in the adjustment of permanent income, following a change in measured income, it seems more plausible than if it is interpreted as reflecting only the costs of adjustment in the money market The current model is not designed to discriminate among the various possible reasons for lags in adjustment [34]

Even with stress given to lags in the formation of expectations, the average lags may appear too long to some It is worth noting, however, that some eminent applied economists have judged adjustment lags to be very long indeed For example, Keynes devoted several pages in the *Treatise* to discussing the effects on European prices of the Spanish gold discoveries of the sixteenth century, quoting Smith's opinion that English prices were first affected in 1570 and that the full effect was not worked out until 1636 [35] Fisher (1930) repeatedly stressed the long lags in the adjustment of price expectations that are implied by the adjustment of nominal interest rates to changes in the rate of inflation The data available to these writers excluded recent experience, when adjustment may be more rapid,[36] but quite a lot of recent econometric evidence, using quarterly postwar data, also suggests that the lags are relatively long The model developed by Bergstrom and Wymer (1974), for example, has average lags of over a year in several equations, and the model as a whole generates an (explosive) 21-year cycle Ando (1974, p 566) reports that the MPS model takes 8–10 years to produce complete "crowding out" of private spending as a result of an increase in government spending, and even longer for the nominal value of output to return to its original level

---

[34] Jonson (1975) presents a second-order model of the monetary sector to illustrate one way in which this problem can be tackled

[35] Keynes (1930, 2 158, n 2)

[36] Estimates of the parameters of the model for the period to 1939, which indicate the effect of postwar experience on the estimates for the whole period, are discussed in Jonson (1975) In general, they are quite similar to those for the full period, an important exception being the estimate of α_7, discussed above

With regard to the structural parameters, perhaps the most interesting finding is the relatively high values for price and interest elasticities. Of particular interest is the price elasticity of demand for exports, which at -3.4 is much higher and more significant than conventional results suggest, and even the impact elasticity ($\alpha_2\beta_6$) is approximately unity As the allocation of changes in expected sales between output and imports is also significantly influenced by relative prices (as indicated by β_7), there seems little doubt that the famous *J*-curve in the U K. current account is indeed a monetary phenomenon [37] The interest elasticities of the demands for money and bonds are also highly significant, and quite large in the latter case. This means in effect that international capital flows respond significantly to changes in domestic and foreign interest rates, and also to anticipated changes in the exchange rate. Interest rate changes also have significant expenditure effects, indirectly via the real balance effect and directly because of the interest effect on desired expenditure, measured by β_3 The net partial equilibrium effect of a sustained change in Bank rate is quite small initially ($\alpha_1\beta_3 + \gamma\beta_{15}$) but increases over time as the money market comes into equilibrium, this is in accordance with the existing evidence about the time profile of interest effects upon expenditures

The two disequilibrium or Phillips effects on price adjustment, represented by β_9 and β_{10}, are significant and have large coefficients, indicating that British prices diverge substantially from purchasing-power parity over a lengthy "short run" during which government spending and private sector domestic spending deviate from their steady-state growth paths This finding, together with the estimate of α_5, suggests that the theoretically attractive small-country assumption of strict commodity price arbitrage must be modified markedly in the British case.

The other interesting results relate to the policy equation, which shows the extent that Bank rate responds systematically to deviations of both income and reserves from their targeted growth paths. There is an interesting asymmetry of response of Bank rate to changes in official loans (which raise both reserves and liabilities) implied by the estimates of β_{22} and β_{23}, which suggests why the British balance of payments has not tended to be stabilized automatically in recent years. The coefficient β_{21} indicates that, if domestic prices rise above traded-goods prices, Bank rate is *reduced* While this seems at first sight to be perverse, it is not entirely irrational if the authorities are aware that a rise in prices above purchasing-power parity will, by itself, tend to produce reductions in prices, and vice versa, as indicated by equations (5) and (6) Their response would then be interpreted as being motivated by an attempt to cushion the expected fall in prices Whatever the reason, this aspect of policy reaction could provide a destabilizing influence on economic fluctuations Nevertheless, calculation of the eigenvalues of the estimated model indicates that the stabilizing feedbacks in the private sector are strong enough to ensure that the system as a whole is stable

---

[37] Cf. Jonson and Kierzkowski (1975) argue, contrary to the conventional wisdom which stresses low or zero price elasticities in the British trade account, that British monetary policy explains the *J*-curve in the U.K. balance of payments following devaluation in 1967.

IV. Some Implications

Figures 11, 12, and 13 present the time paths of the price level, output, and international reserves over the sample period, along with the control solution values (represented by the dotted lines) obtained from a dynamic simulation of the model from 1885 A dynamic simulation of a macroeconometric model is in effect a very stringent test of its goodness of fit, and the figures illustrate that the model tracks broad historical experience rather well A slightly modified version of the model[38] also tracks quite well the experience beyond the sample period from 1971 to 1974, and in particular explains the recent takeoff of British inflation, which rose from an annual rate of approximately 7 percent in 1970 to almost 12 percent in 1974 Given that the model explains British macroeconomic experience, it can be used to analyze the contributions of each exogenous variable to deviations of the endogenous variables from their steady-state growth paths The following account discusses postwar experience; this is compared with the earlier experience by Jonson (1975), who presents the details of the simulation analysis discussed here

Britain, along with her major trading partners, entered the postwar period with enormous excess money balances, and a further major monetary shock was provided by devaluation in 1949, which substantially changed the relative prices of traded and nontraded goods during the adjustment period These disturbances together resulted in a period of excess demand for British goods, which took a considerable period to work itself through the system, if the estimates of the lags presented above are approximately correct [39] Thus, the historically low levels of unemployment in the 1950s and early 1960s were, according to the present model, in part a consequence of two major monetary disturbances[40] as well as the increased fiscal activism which has been practiced in postwar Britain It is conceivable that the effects of two major monetary shocks convinced

---

[38] The modification was introduced to allow for the Heath-Barber tax cuts in the early 1970s and the attempt to establish "competition and credit control" in the banking sector

[39] Friedman (1972) also conjectured, on the basis of less formal evidence, that U K money balances took a considerable period to return to normal following World War II

[40] Compare Johnson's (1975) suggestion that the revaluation and monetary stringency of the early 1920s raised the unemployment rate above its equilibrium or natural level for a considerable period and produced the conditions which motivated Keynes's *General Theory*.

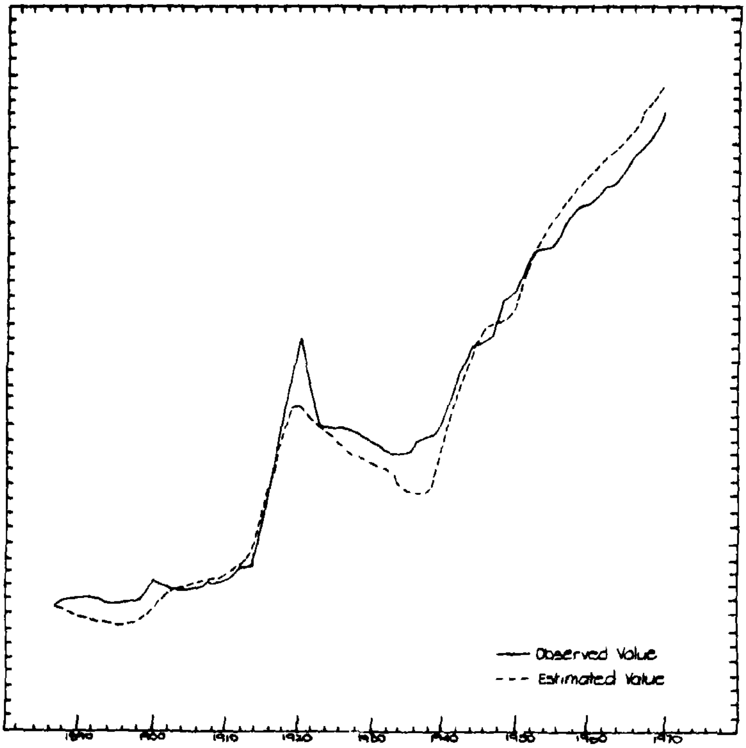


Fig. 11.—Prices actual and control solution

British policymakers that they had solved the unemployment problem solely by using Keynesian demand management policies. When the effect of the monetary disturbances began to wear off, it would then seem appropriate to resort to more and more fiscal activism, driving the price level above purchasing-power parity and necessitating the 1967 and more recent devaluations.

The postwar period has seen relatively strong growth in world income and a rising world price level. By itself, the buoyancy of world demand and prices accounts for most of the postwar rise in British prices and, together with the large official loans, has enabled U.K. international reserves to grow slowly. The simulations indicate, however, that if real U.K. government spending had grown at the same rate as real income, international reserves would have grown much more rapidly, and the United Kingdom would not have had to resort to borrowing to maintain its reserves. In this sense, it can be said that the rest of the world has financed much of the postwar growth of the U.K. government sector.

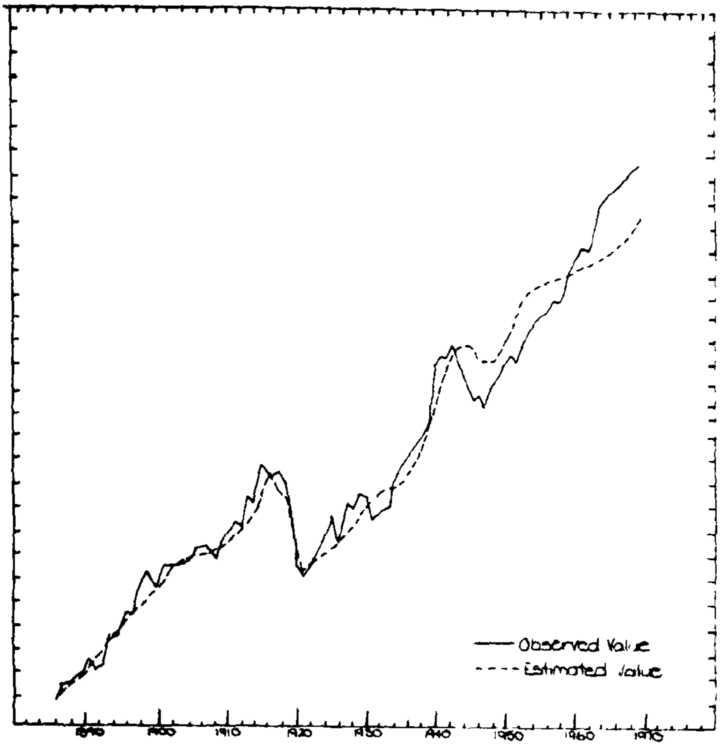


Fig 12 —Output actual and control solution

This strong growth of government spending, with the associated monetary effects, explains both the tendency for British prices to rise more rapidly than world prices and the slow growth of Britain's international reserves These results suggest in particular that the current British macroeconomic crisis is a direct result of the Heath-Barber dash for growth in the early 1970s and the accelerated growth of government spending more recently The model suggests that there is little mystery about current British inflation and Britain's associated macroeconomic problems, and no need to resort to untestable assertions about union leaders "plucking numbers from the air"[41] Return to a more acceptable configuration of price, output, and reserves growth will involve a painful period of relatively slower growth of demand, particularly by the government sector The costs of this have to be compared with those of the continued or accelerating inflation which will result if recent trends persist.

---

[41] Wiles (1973).

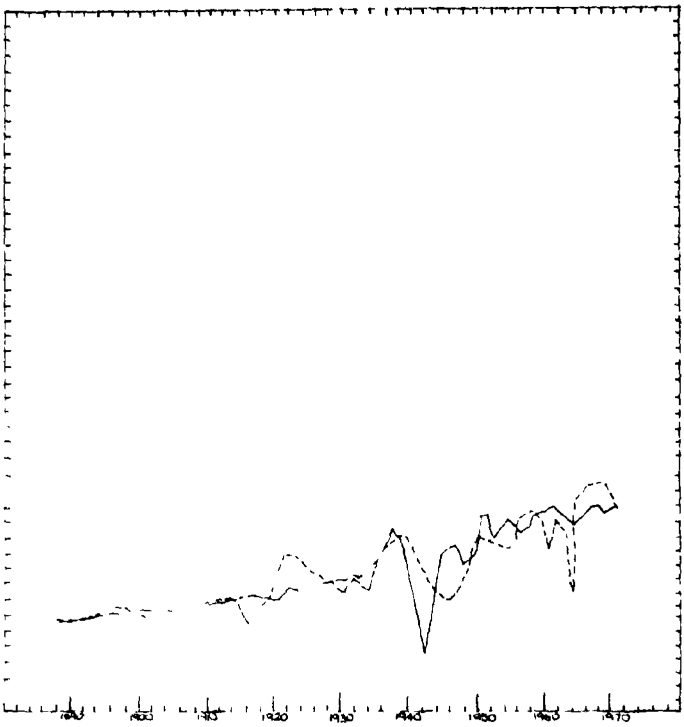


FIG 13.—Official reserves actual and control solution

Appendix A

The Model as a Whole

$$\begin{aligned} D \log d &= \alpha_1(\log \hat{d} - \log d) + \gamma(\log \hat{m} - \log M/p) \\ \hat{d} &= d_0 e^{\beta_1 Q2}(y - T/p)^{\beta_2}(r_b)^{\beta_3} \end{aligned} \tag{1}$$

$$\begin{aligned} D \log x &= \alpha_2(\log \hat{x} - \log x) + \beta_4 \log (y/y^* e^{\lambda_1 t}) \\ \hat{x} &= x_0 e^{\beta_5 Q2}(y_w)^{\beta_2}(P_x/P_l)^{\beta_6} \end{aligned} \tag{2}$$

$$\begin{aligned} D \log \imath &= \alpha_3(\log \hat{\imath} - \log \imath) \\ \hat{\imath} &= \imath_0(P_l/P)^{\beta_7} S^e \end{aligned} \tag{3}$$

$$\begin{aligned} D \log y &= \alpha_4(\log \hat{y} - \log y) \\ \hat{y} &= [1 - \imath_0(P_l/P)^{\beta_7}]S^e \\ S^e &= (d + x + g)e^{\beta_8 Q1} \end{aligned} \tag{4}$$

$$\begin{aligned} D \log P = \alpha_5[\log EP_w - \log P + \beta_9 \log (d/d^* e^{\lambda_1 t}) \\ + \beta_{10} \log (g/g^* e^{\lambda_1 t})] + \beta_{11} Q2^* \end{aligned} \tag{5}$$

$$D \log P_x = \alpha_6(\log P_l - \log P_x) + \beta_{12} D \log P_l + \beta_{13} D \log P \tag{6}$$

$$\begin{aligned} D\log M &= \alpha_7(\log \hat{m} - \log M/p) \\ \hat{m} &= m_0 e^{\beta_{14}Q1} y^{\beta_2} r_b^{\beta_{15}} (r_w + F)^{\beta_{15}} \end{aligned} \tag{7}$$

$$\begin{aligned} D\log B &= \alpha_8(\log \hat{b} - \log B/p) \\ \hat{b} &= b_0 e^{\beta_{16}Q1'} y^{\beta_2} r_b^{\beta_{17}} (r_w + F)^{-\beta_{17}} \end{aligned} \tag{8}$$

$$\begin{aligned} D\log r_b = {} & \alpha_9 \log (r_0/r_b) + \beta_{18} D\log r_w + \beta_{19}Q3 \\ & + \beta_{20}\log (y/y^{*}e^{\lambda_1 t}) + \beta_{21}\log (P/P_t) \\ & + \beta_{22}\log [R/R^{*}e^{(\lambda_1+\lambda_2)t}] + \beta_{23}\log [L/L^{*}e^{(\lambda_1+\lambda_2)t}] \end{aligned} \tag{9}$$

$$T = T_0 e^{\beta_{24}t}(Py)^{\beta_2} \tag{10}$$

$$DC = Pg - T - DB \tag{11}$$

$$DR = DM - DC + DL \tag{12}$$

Appendix B

The endogenous variables are

d = real expenditure (consumption plus investment)
x = real exports
i = real imports
y = output
P = price of output
P_x = price of exports
M = money balances
B = net bonds
r_b = bank rate
T = tax receipts
C = domestic credit
R = sterling value of official reserves

The exogenous variables are

g = real U K government current spending
y_w = world income (proxied by U S. GNP)
P_w = world prices (proxied by U S. GNP deflator)
P_t = world traded-goods prices in sterling (proxied by U.K. import prices)
r_w = world interest rates (proxied by U S. bank rate)
E = exchange rate (proxied by inverse of dollar-pound exchange rate)
F = expected change in the exchange rate (proxied by the forward discount)
L = official loans and cumulated value of official revaluation gains and losses
$Q1$ = dummy variable with value unity for 1917–21, zero otherwise
$Q1'$ = dummy variable with value unity for 1916–20, zero otherwise
$Q2$ = dummy variable with value unity for 1940–47, zero otherwise
$Q3$ = dummy variable with value unity for 1940–1951, zero otherwise
t = time

The expected coefficient signs are

$\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9 > 0, \gamma < 0$
$\beta_2, \beta_4, \beta_9, \beta_{10}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{16}, \beta_{17}, \beta_{18}, \beta_{20}, \beta_{23}, \beta_{24} > 0$
$\beta_1, \beta_3, \beta_5, \beta_6, \beta_7, \beta_8, \beta_{11}, \beta_{15}, \beta_{19}, \beta_{22} < 0$
$\beta_{21} \lesseqgtr 0$

(The steady-state levels of the targeted values in the policy reaction function cannot be independently estimated, and y^*, R^*, and L^* are set equal to the values of output, reserves, and loans in 1884. The only result of an incorrect choice of these variables is to bias the estimated constant in the policy reaction function (r_0). The coefficients whose signs are indeterminate a priori are the constants in various equations, except for β_{21}, the relative price elasticity in the policy reaction function, discussed above.)

References

Anderson, C. J.; Pearce, I. F., and Trivedi, P. K. "Output, Expected Demand and Unplanned Stocks." Paper presented to the Association of University Teachers of Economics Conference, Sheffield, April 1975.

Ando, A. "Some Aspects of Stabilization Policies: The Monetarist Controversy and the M.P S. Model." *Internat. Econ. Rev* 15, no. 3 (October 1974): 541–71

Archibald, G. C , and Lipsey, R. G. "Monetary and Value Theory. A Critique of Lange and Patinkin." *Rev. Econ Studies* 26, no. 69 (October 1958) · 1–22.

Ball, R. J., Boatwright, B., Burns, T., Lobban, P., and Miller, G. *The London Business School Quarterly Econometric Model of the U K. Economy.* Econometric Forecasting Unit Discussion Paper no. 20. London London Bus. School, 1973.

Ball, R. J., Eaton, J. R.; and Steuer, M. D. "The Relationship between U.K. Export Performance in Manufactures and the Internal Pressure of Demand." *Econ. J.* 76 (September 1966) 501–18.

Bergstrom, A. R., and Wymer, C. R. *A Model of Disequilibrium Neoclassical Growth and Its Application to the United Kingdom* International Monetary Research Project Discussion Paper. London. London School Econ., September 1974.

Brunner, K , and Meltzer, A. H. "Monetary and Fiscal Policy in Open Interdependent Economies with Fixed Exchange Rates." Paper presented at third Paris-Dauphine Conference, March 1974

Cagan, P *Determinants and Effects of Changes in the Stock of Money, 1875–1960.* New York Columbia Univ. Press, 1965.

Carlson, J. A., and Parkin, J. M. "Inflation Expectations." *Economica* 42, no 166 (May 1975) 123–38.

Courchene, T , and Singh, K. "The Monetary Approach to the Balance o Payments An Empirical Analysis of 14 Industrial Countries." In *Inflation as a World Phenomenon*, edited by J M. Parkin and G Zis. Manchester Manchester Univ Press, 1975.

Davis, R. G. "How Much Does Money Matter? A Look at Some Recent Evidence." *Federal Reserve Bank New York Monthly Rev.* 51, no. 6 (June 1969): 119–31.

Dornbusch, R., and Mussa, M. "Consumption, Real Balances and the Hoarding Function." *Internat Econ. Rev.* 16, no 2 (June 1975): 415–21.

Fisher, G., and Sheppard, D. *Effects of Monetary Policy on the United States Economy: A Survey of Econometric Evidence.* OECD Economic Outlook Occasional Studies. Paris: Org. Econ Cooperation and Development, December 1972.

Fisher, I. "Will the Present Upward Trend of World Prices Continue?" *A.E.R.* 2 (September 1912). 531–58.

———. *The Theory of Interest.* New York: Macmillan, 1930.

Frenkel, J. A., and Rodriguez, C. A. "Portfolio Equilibrium and the Balance of Payments: A Monetary Approach." *A.E.R.* 65 (September 1975): 674–88.

Friedman, M. "Monetary Trends in the United States and the United Kingdom." *American Econ.* 16, no. 1 (Spring 1972): 4–17.

Friedman, M., and Meiselman, D. "The Relative Stability of Monetary Velocity and the Investment Multiplier in the U.S. 1879–1958." In *Stabilization Policies*, prepared for the Commission on Money and Credit by E. C. Brown et al. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

———. "Money and Business Cycles." *Rev. Econ. and Statis.* 45, suppl. (February 1963): 32–78. Reprinted in "*The Optimum Quantity of Money" and Other Essays.* Chicago: Aldine, 1969.

Friedman, M., and Schwartz, A. J. *A Monetary History of the United States, 1867–1960.* Princeton, N.J.: Princeton Univ. Press, 1963.

Genberg, H. A. "Aspects of the Monetary Approach to Balance of Payments Theory: An Empirical Study of Sweden." In *The Monetary Approach to the Balance of Payments*, edited by H. G. Johnson and J. Frenkel. London: Allen & Unwin, 1975.

Johnson, H. G. "Keynes and British Economics." In *On Economics and Society.* Chicago: Univ. Chicago Press, 1975.

Johnson, H. G., and Frenkel, J., eds. *The Monetary Approach to the Balance of Payments.* London: Allen & Unwin, 1975.

Jonson, P. D. "Stabilization Policy in Australia: An Objective Analysis." *Manchester School* 42, no. 3 (September 1974): 259–76.

———. "An Investigation of the U.K. Balance of Payments with Particular Emphasis on the Role of Monetary Factors and Disequilibrium Dynamics, 1882–1970." Ph.D. dissertation, London Univ., 1975.

Jonson, P. D., and Kierzkowski, H. I. "The Balance of Payments: An Analytic Exercise." *Manchester School* 43, no. 2 (June 1975): 105–33.

Keynes, J. M. *A Treatise on Money.* London: Macmillan, 1930.

Knight, M., and Wymer, C. R. *A Monetary Model of an Open Economy, with Particular Reference to the United Kingdom.* International Monetary Research Project Discussion Paper. London: London School Econ., 1975.

Kouri, P., and Porter, M. "International Capital Flows and Portfolio Equilibrium." *J.P.E.* 82, no. 3 (May/June 1974): 443–67.

Laidler, D. E. W. "Price and Output Fluctuations in an Open Economy." In *Essays on Money and Inflation*, by D. E. W. Laidler. Manchester and Chicago: Univ. Manchester Press and Univ. Chicago Press, 1975.

Lucas, R. E. "Some International Evidence on Output-Inflation Trade-offs." *A.E.R.* 63, no. 3 (June 1973): 326–34.

Moroney, J. R., and Mason, J. M. "The Dynamic Impacts of Autonomous Expenditures and the Monetary Base on Aggregate Income." *J. Money, Credit, and Banking* 3, no. 4 (November 1971): 793–814.

Mussa, M. "A Monetary Approach to Balance-of-Payments Analysis." *J. Money, Credit, and Banking* 6, no. 3 (August 1974): 333–51.

Norton, W. E. "Debt Management and Monetary Policy in the United Kingdom." *Econ. J.* 79 (September 1969): 475–94.

Norton, W. E., and Henderson, J. F. *A Model of the Australian Economy. A Further Report.* Reserve Bank of Australia Occasional Paper no. 3G. Sydney: Reserve Bank Australia, March 1972.

Parkin, J. M. "Inflation, the Balance of Payments, Domestic Credit Expansion, and Exchange Rate Adjustments." Paper presented to the first Racine Conference, Racine, Wis., November 1972.

———. "World Inflation, International Relative Prices and Monetary Equilibrium under Fixed Exchange Rates." Paper presented to second Racine Conference, Racine, Wis., July 1974.

Parkin, J. M.; Henderson, J. F.; Cooper, R. L.; and Danes, M. K. "Portfolio Behaviour: Choices between Real and Financial Assets." Paper presented to the forty-fifth conference of the Australian and New Zealand Association for the Advancement of Science, Perth, August 1973.

Patinkin, D. *Money, Interest and Prices*. 1st ed., Evanston, Ill.. Row Peterson, 1956; 2d ed., New York: Harper & Row, 1965.

Phelps, E. S., et al *The Microeconomic Foundations of Employment and Inflation Theory*. New York: Norton, 1970.

Sargent, T. J. "A Note on the Accelerationist Controversy " *J. Money, Credit, and Banking* 3, no 3 (August 1971): 721–25.

Wiles, P. J. "Cost Inflation and the State of Economic Theory." *Econ J* 83 (June 1973) 377–98.

Wymer, C. R. "Econometric Estimation of Stochastic Differential Equation Systems." *Econometrica* 40, no 3 (May 1972): 565–77.

Private Pensions and Saving: New Evidence

Alicia H. Munnell

Federal Reserve Bank of Boston

This paper examines the impact of private pension coverage on the saving behavior of men in their preretirement years The empirical work is based on the Ando-Modigliani model which permits explicit recognition of differences in expected retirement age between covered and noncovered groups The data originated in the 5-year Labor Department sample of men aged 45–59 in 1966 The results clearly indicate that, contrary to earlier work by Cagan and Katona, pension coverage reduces saving in other forms

With the enormous growth of both Social Security and private pensions (see table 1), there has been increasing concern about the impact of these retirement programs on individual saving behavior The response of individuals to guaranteed retirement income is important because the future welfare of retired persons depends significantly on whether they reduce their own saving for each dollar of expected pension benefit or whether they use their pensions as a base on which to build for a truly secure retirement Furthermore, in order to predict the future supply of capital it is essential to know whether the reserves accumulated in pension funds are being offset by lower individual saving or whether these reserves represent a net addition to the supply of capital

This paper will present the results of a study on the effect of pension coverage on saving behavior for a sample of approximately 5,000 men who were between the ages of 45 and 59 in 1966 Data were collected by the Bureau of the Census, under contract from the Labor Department, in a series of surveys over the 5-year period 1966–71 Contrary to earlier empirical work on the effect of pension coverage on saving, the findings

I would like to thank Martin S Feldstein for helpful comments on an earlier draft of this paper. I am also totally indebted to Lucy Rexroad and Joanne Grolnic for their computer support. An earlier draft of this paper was presented at the National Bureau of Economic Research Conference on "The Economic Impact of Pensions" in Washington, D.C., May 1975

[*Journal of Political Economy*, 1976, vol 84, no 5]
© 1976 by The University of Chicago All rights reserved

TABLE 1

GROWTH OF PRIVATE PENSIONS, 1950–73

	CONTRIBUTIONS ($ millions)			BENEFITS		
	Total	Employer	Employee	*N* Beneficiaries (thousands)	Amount of Payments ($ millions)	RESERVES, END OF YEAR ($ billions)
1950	2,080	1,750	330	450	370	12.1
1955	3,840	3,280	560	980	850	27.5
1960	5,490	4,710	780	1,780	1,720	52.0
1965	8,360	7,370	990	2,750	3,520	86.5
1970	14,000	12,580	1,420	4,720	7,360	137.1
1973	21,105	19,390	1,715	6,130	11,220	180.2

SOURCE —Kolodrubetz (1975)

of this study indicate clearly that pension coverage reduces saving efforts in other areas—at least for those older men for whom retirement is the primary saving motive

Part I of this paper will discuss the theoretical alternatives and comment briefly on previous work on pensions and saving. Part II describes the nature of the sample used in this study, while Part III presents the empirical results The final section provides the conclusions and their implications for aggregate saving

I. The Effect of Pensions in a Life-Cycle Saving Model

Pension coverage could affect saving behavior in either of two ways Traditional saving theory indicates that private saving would decline as individuals respond to the additional security provided by pension coverage However, the more popular view, based on two cross-sectional studies by Cagan (1965) and Katona (1965), has been that the prospect of private pension benefits stimulates saving by providing a base on which to build toward an adequate retirement income.

The study by Phillip Cagan (1965) analyzed the saving behavior in 1958–59 of over 15,000 members of the Consumers Union Cagan's explanation of the surprising results was that pension coverage calls attention to retirement needs and prospects, and thereby fosters a "recognition effect" that counteracts individuals' disinclinations to plan for the future [1]

---

[1] As part of a study on *The Effect of Social Security on Personal Savings* (1974), this author reanalyzed a subsample of Cagan's Consumer Union survey using multiple regression techniques and the results directly contradicted Cagan's early conclusions There seem to be three contributory factors that can explain the contradictory results First, the subsample contains more information than Cagan's since it includes all four rounds of questionnaires Furthermore, the savings data are constructed through detailed balance sheet information rather than by respondents' estimates of "approximate changes" in broad categories of assets The subsample was also subjected to considerably more elaborate screening and consistency checks. The second factor contributing to the new results is the method of analysis. Owing to the size of Cagan's sample, he was not able to standardize for a large number of socioeconomic variables simultaneously. Multiple regression analysis made it possible to isolate the impact of pension coverage from that of

The study by George Katona (1965), based on personal interviews with about 2,000 consumers in 1962–63, also concluded that pension plans stimulate voluntary saving Katona added a second explanation for his results, hypothesizing a "goal feasibility" effect where individuals intensify their saving efforts the closer they get to their retirement goal This explanation makes saving preferences endogenous and constitutes, therefore, a dramatic departure from traditional consumer theory

Recent studies of the impact of Social Security benefits on personal saving tend to contradict the private pension findings and are more consistent with life-cycle saving theory—see, for example, Munnell (1974) and Feldstein (1974) These studies suggest that individuals have reduced their saving in response to their guaranteed benefits under the Social Security system These studies have also brought out a secondary impact of pension plans on saving these plans might influence saving behavior indirectly by altering the retirement decision In the case of Social Security, it seemed very clear that the declining labor force activity of the aged since the introduction of the system has strongly stimulated saving, thereby weakening the overall depressant effect of guaranteed retirement benefits

This retirement effect may have contributed to the surprising results in the early studies that pension coverage stimulated saving. Since pension coverage is usually accompanied by compulsory retirement,[2] covered employees may simply have increased their saving in anticipation of an involuntary retirement Fortunately, a fully specified life-cycle saving model can make allowance for differences in expected retirement age between those covered and those not covered by pension plans

The framework for analysis in this study was the Ando-Brumberg-Modigliani life-cycle saving model where consumption is a function of lifetime resources—see Modigliani and Brumberg (1954) and Ando and Modigliani (1963)

$$C_t = \frac{1}{LE}[Y_t + (D - T)Y^e + A_{t-1}] \qquad (1)$$

---

education, family size, income, age, and other factors Finally, the analysis of the subsample focused on the saving behavior of those persons most concerned about saving for retirement, individuals aged 55–64 If a substitution effect of pension saving for private saving existed it would clearly be found among members of this preretirement age group Those interested in further details may contact the author for a summary of this analysis

[2] Evidence that private pension plans and retirement go together was provided by a study of compulsory retirement among firms employing 50 or more workers, which "confirmed the fact that plants which have neither a pension nor a profit-sharing plan tend overwhelmingly to have a completely flexible retirement policy," (Slavick 1966, p. 16). In 1963, a Bureau of Labor Statistics study confirmed the widespread coexistence of pension plans and compulsory retirement. At that time, nearly 75 percent of private plans with more than 60 percent of covered workers had involuntary retirement provisions (U.S. Department of Labor 1965, p 27)

where D = age at death = $(T + LE)$, T = age in period t, Y_t = actual income in period t, Y^e = expected income in subsequent periods, A_{t-1} = asset holdings at beginning of period t, and LE = life expectancy. The expected future income expression can be rewritten as the sum of future earned income plus expected pension benefits after retirement Assuming future earned income is equal to income in period t (or that the rate of growth of earnings is equal to the discount rate), then total expected income can be written as follows.

$$(D - T)Y^e = (R - T)Y_t + (D - R)\,PEN, \tag{2}$$

where R = expected retirement age

Therefore,

$$C_t = \frac{1}{LE}[Y_t(1 + R - T) + (D - R)\,PEN + A_{t-1}] \tag{3}$$

$$S_t = Y_t - C_t \tag{4}$$

$$S_t = Y_t - \frac{1}{LE}[Y_t(1 + R - T) + (D - R)\,PEN + A_{t-1}] \tag{5}$$

$$= \frac{(D - R - 1)}{LE}Y_t - \frac{(D - R)}{LE}PEN - \frac{A_{t-1}}{LE} \tag{6}$$

where PEN = expected pension benefits All the empirical work was carried out within the framework of equation (6), although other variables were added to help standardize for variation in socioeconomic characteristics

II. Data

The data for this study were derived from a series of surveys conducted by the Bureau of the Census, under contract from the Department of Labor, over a 5-year period from 1966 to 1971 The subsample used in the paper included employed male heads-of-households who were aged 45–59 in 1966 Although the main purpose of the surveys was to "describe and explain labor force participation, employment status and mobility," considerable data were also collected on financial and other characteristics, including whether or not persons expected to receive pension benefits in addition to Social Security and Railroad Retirement.[3]

The Labor Department sample is fairly representative of the population in general. The only deviation is considerable oversampling of black men in order to provide statistically reliable estimates for this group. The sample cases are weighted to reflect the different sampling ratios for whites and blacks so that the sample can be adjusted to reflect the entire noninstitutional population. In this study, however, consistent data were available for only about 50–60 percent of the original sample, so that all results are presented in terms of the actual number of cases rather than the entire population

---

[3] A detailed description of the survey is found in U.S. Department of Labor, Manpower Administration (1970, 1973).

A comparison of the income distribution and educational attainment of the Labor Department sample with census data for the relevant age group published regularly in the *Current Population Reports* reveals that the sample contains a somewhat larger percentage of low-income individuals than exists in the male preretirement population. In terms of educational attainment, the sample has a slightly lower percentage with high school and college education than the entire U S age group [4]

The specific explanatory variables used in the empirical analysis include income and wealth data, saving information defined in two alternative fashions, dummy variables for pension coverage, college education, race, desire to leave inheritance, self-employment, and finally age and life expectancy data Since asset data were only available for the years 1966, 1969, and 1971, saving behavior was analyzed over two time periods 1966–69 and 1969–71

Income —Net family income was available for each year, 1965–70, and was one of the key variables recreated by the Labor Department (on the basis of detailed decision rules) to be as consistent as possible with counterpart variables in other years In accordance with the model, 1966 income was used for the 1966–69 period, while 1969 income was used for the 1969–71 analysis In addition, a permanent income variable was constructed by calculating a simple arithmetic average for each respondent based on the number of years for which income was reported [5]

Wealth —Detailed information on family assets was collected in 1966, 1969, and 1971 This net family asset variable was also recreated by the Labor Department to be consistent with asset data in other periods and income data within the period For 1966 and 1969, there are no automobile asset data in the wealth series, consequently, liabilities on autos also were excluded The 1971 net family asset figure had to be adjusted to exclude both auto asset and liability information to yield consistent data.

Saving.—Two definitions of saving were computed for the analysis The first saving data were constructed by calculating the annual change in net family assets for the periods 1969–66 and 1971–69. The advantage of these data was their reliability since net family assets had been carefully checked for consistency. The major disadvantage of these saving data, however, was the inclusion of capital gains—particularly through appreciation of house values[6] and, less importantly, through changes in the market value of securities.

---

[4] Income distribution for census sample from *Current Population Reports* Series P-60, no 80, "Consumer Income," p 34 Family income with male head, aged 55–64, all races Educational attainment distribution for census sample from *Current Population Reports* Series P-20, no 274, "Educational Attainment in the U S , March 1973 and 1974 " Years of education for males, age 55 years and over, all races

[5] Only 1,506 of the original 5,030 respondents reported income in all tax years Therefore, the technique adopted was to calculate an arithmetic average for each respondent based on the available data. To construct a more complex measure of permanent income (such as one with geometrically declining weights) would have resulted in an inordinate reduction in the sample size.

In an attempt to find data that would more closely approximate saving out of current income, an alternative saving variable was constructed based on changes in the original components of the net family asset data. With this approach, it was possible to eliminate capital gains through rising house prices, although gains through the stock market remained. Both saving variables were used in the empirical work in order to test the sensitivity of the results to the particular definition of saving.

Retirement age.—A series of questions was asked concerning compulsory retirement age, whether the respondent planned to retire early or (in the absence of compulsory retirement) the age at which he planned to retire. From this set of questions an "expected retirement age" variable was constructed. For the purposes of this study, those persons who failed to answer these questions were assigned an expected retirement age of 65.

Life expectancy.—Each respondent was assigned a life expectancy for each of the three years (1966, 1969, and 1971) based on his age in the particular year and on his race (U.S. Department of Health, Education, and Welfare 1974, sec. 5, table 5-3).

Pension coverage.—Coverage by a private pension plan or Social Security were each represented by a dummy variable set equal to 1 if individuals expected to receive either private pension and/or Social Security benefits.

Socioeconomic variables.—Some socioeconomic variables were included in the regressions in an attempt to specify the saving relationship more precisely. The four dummy variables reflect race, education, attitude toward inheritance, and employment status: *COLL* = 1 (respondent experienced at least some college education), *INHER* = 1 (respondent wants to leave inheritance), *RACE* = 1 (respondent is nonwhite), *SEMP* = 1 (respondent is self-employed).

---

[6] The upward trend in house prices was definitely reflected in the survey information as demonstrated in the following table.

PERCENTAGE DISTRIBUTION OF HOUSE VALUES, 1966–71

House Value ($)	1966	1969	1971
1–9,999	27.4	23.4	19.6
10,000–14,999	20.8	18.5	15.9
15,000–24,999	32.2	30.1	29.6
25,000–49,999	17.7	23.5	29.5
50,000+	1.9	4.5	5.4
Total	100.0	100.0	100.0

Complete data were available for only 2,600 of the original 5,030 cases. To guard against using a biased subgroup of observations, the means of all important variables to the subsample were checked against those for the entire survey Fortunately, there were no significant differences in the mean values of the relevant variables in the two groups

III. Empirical Results

As a prelude to exploring saving behavior, average net worth and net financial assets were classified by age and income in order to compare the wealth of individuals expecting pension benefits with those not covered by private retirement plans [7] The results of these tabulations (example summary in table 2) for 1966, 1969, and 1971 reveal that, except for very low-income individuals, those men not covered by pensions fairly consistently had greater wealth than those expecting benefits [8]

To provide additional support for the hypothesis that persons covered by pensions save less than those not covered, the life-cycle saving model developed in the first section was tested empirically Several variants of the original formulation were tried for the two periods 1966–69 and 1969–71 and the results are summarized in tables 3–5 [9] The results consistently support the hypothesis that pension coverage reduces saving in other forms

According to the model presented in the first section, saving is a function of the following variables

$$S_t = \frac{(D - R - 1)}{LE} Y_t - \frac{(D - R)}{LE} PEN - \frac{A_{t-1}}{LE} \qquad (6)$$

---

[7] Coverage under pension plans includes workers who are covered by the civil service and military pensions as well as workers with private industry plans In terms of the certainty of receiving benefits, the government plans are more like Social Security On the other hand, government pensions are partially funded in contrast to Social Security's pay-as-you-go financing

[8] Similar evidence has already been presented for 1966 by Gold (1972)

[9] The 1966–69 sample included 2,649 observations, while the 1969–71 originally included 2,608 However, five outliers were eliminated from the later sample because they distorted the positive relationship between income and saving This appeared to be legitimate in view of the high capital losses reported by the outliers, all of whom had incomes in excess of $30,000 but dissaving in excess of $20,000 When the equations were then fitted to the smaller (2,603) sample, the coefficients conformed to those derived from the earlier sample period The outlying observations reported the following

Income	Wealth	Saving
$40,000	$655,000	–$188,000
50,000	220,000	–17,500
32,000	207,000	–60,500
39,000	52,000	–19,500
31,000	255,000	–69,000

TABLE 2

1969

Joint Income Bracket ($)	Expecting Benefits				Not Expecting Benefits			
	Respondents		Average Net Worth* ($)	Average Net Liquid Assets† ($)	Respondents		Average Net Worth* ($)	Average Net Liquid Assets† ($)
	N	% Total			*N*	% Total		
Age 48–52:								
Under 2,500	16	2 3	27,373	1,577	56	10 4	4,469	−[illegible]
2,500 under 5,000	30	4 3	4,065	45	112	20 9	8,690	[illegible]
5,000 under 7,500	90	13 0	8,705	808	116	21 6	11,803	[illegible]
7,500 under 10,000	125	18 0	11,417	1,512	100	18.7	17,322	2,[illegible]
10,000 under 15,000	231	33 4	19,988	3,625	111	20.7	26,397	1,[illegible]
15,000 and over	201	29 0	46,276	15,508	41	7.6	101,429	19,171
Total‡	693	100 0	24,083	6,122	536	100.0	21,294	2,312
Age 53–57:								
Under 2,500	8	1 5	3,586	−573	103	18.3	7,003	622
2,500 under 5,000	27	4.9	7,823	847	128	22.7	11,561	441
5,000 under 7,500	101	18 2	11,531	1,211	114	20.2	15,805	1,995
7,500 under 10,000	113	20 4	20,354	2,756	92	16.3	18,000	3,689
10,000 under 15,000	179	32.3	22,619	5,287	81	14 4	36,051	7,180
15,000 and over	126	22 7	74,823	19,319	45	8 0	112,627	19,567
Total‡	554	100 0	31,013	6,918	563	100 0	24,240	3,818

Age 58–62:								
Under 2,500	14	3.6	12,337	2,047	121	24.6	12,556	1,346
2,500 under 5,000	33	8.6	10,014	2,287	129	26.2	9,116	1,027
5,000 under 7,500	70	18.2	11,984	2,062	88	17.9	16,234	2,687
7,500 under 10,000	94	24.4	20,389	6,804	48	9.8	20,334	3,245
10,000 under 15,000	94	24.4	33,586	9,724	65	13.2	24,364	6,151
15,000 and over	80	20.8	83,927	30,200	41	8.3	141,283	40,224
Total‡	385	100.0	34,104	10,956	492	100.0	25,358	5,562
Age 48–62:								
Under 2,500	38	2.3	16,826	1,297	280	17.6	8,896	735
2,500 under 5,000	90	5.5	7,374	1,108	369	23.2	9,835	755
5,000 under 7,500	261	16.0	10,678	1,300	318	20.0	14,464	1,595
7,500 under 10,000	332	20.3	16,999	3,434	240	15.1	18,185	3,041
10,000 under 15,000	504	30.9	23,459	5,353	257	16.2	28,925	4,650
15,000 and over	407	24.9	62,514	19,576	127	8.0	118,263	26,108
Total‡	1,632	100.0	28,799	7,553	1,591	100.0	23,593	3,917

---

* Average net worth includes the value of residential and business assets less mortgage debt, savings accounts, stocks, bonds, loans to others less other debts. Automotive asset and liabilities are excluded

† Net liquid assets is equal to average net worth less value of assets and liabilities associated with residence and business.

‡ Since totals are weighted averages, those for persons expecting benefits will exceed those for persons not expecting benefits because pension coverage is positively related to income.

From this theory, the propensity to save out of earned income, Y_t, is proportional to the ratio of an individual's retirement span $(D - R - 1)$ to his total life expectancy (LE) Total savings will be reduced by a fraction, also dependent on retirement and/or life expectancy, of expected pension income and total assets. To test this model, each individual's earned income and expected pension income were introduced into the equations weighted by the appropriate ratio of retirement years to life expectancy and assets were introduced divided by the individual's life expectancy.

An example will help to clarify the nature of the equations. The average individual in this sample was 51 years old in 1966, with a life expectancy of 22 more years and an expected retirement age of 65, and had an income of \$8,350 and total assets of \$19,500 The income variable entered into the regression for this individual would be $[(73 - 65 - 1)/22] \times \$8,350 = \$2,672$ and the asset variable would be $\$19,500/22 = \$886$

In short, each income and asset variable in the equation is weighted by the theoretical propensity to save. If the theory is correct, then the coefficients of income and assets each weighted by the hypothetical saving propensities should be equal to $+1$ and -1, respectively.[10] If data on expected pension income had been available, then the coefficient of this item weighted by the ratio of retirement years to life expectancy should also be -1 However, since dummy and proxy variables were used for pension income, the expected coefficients for the pension variables have to be discussed within the context of the specific models presented below

The dependent variable in the following equations is the dollar amount of saving based on the annual change in net family assets for the periods 1966–69 and 1969–71 With this definition, annual saving averaged \$1,904 for 1966–69 and \$2,275 for the later period This saving variable was used because the extensive consistency checks made possible greater confidence in the accuracy of the information. All equations were reestimated for the alternative saving definition which excluded capital gains on houses, and the results were consistent with the equations in tables 3–5.

Model 1 —The first formulation was based on equation (6), but in the absence of explicit expected pension benefits, dummy variables were introduced for coverage by private pension plans and Social Security. Therefore, the first set of regressions was estimated using equation (7)

$$\begin{aligned} S_t = \alpha_1[(D - R - 1)/LE]\, Y_t - \alpha_2[(D - R)/LE] \cdot PP \\ - \alpha_3[(D - R)/LE] \cdot SS - \alpha_4[(A_{t-1})/LE], \end{aligned} \tag{7}$$

where SS and PP indicate coverage under Social Security and private plans, respectively. In keeping with the theoretical framework, the dummy variables as well as the income and asset variables were multiplied by the propensity to save.

---

[10] Actual coefficients would deviate somewhat from the hypothetical values of unity if the current income used in the regression were lower or higher than a household's expected lifetime income. Presumably current income would be below lifetime income for younger households and above that for older households. Therefore, insofar as consumer durables including houses are excluded from saving, the coefficients would be lower in the saving equation for younger households and higher for older households.

The equations are presented without an intercept since the theory implies that the function should pass through the origin.[11] As discussed above, the coefficients of the weighted income and asset terms should be +1 and −1. The coefficients of both the Social Security and pension dummies should be negative, while all four socioeconomic variables—education, desire to leave inheritance, self-employment, and race—would be expected to have a positive effect on saving.[12]

The results obtained by estimating equation (7) for 1966–69 and 1969–71 are presented in table 3. All coefficients are reasonable in sign and magnitude. Wealth and income have coefficients amazingly close to −1 and +1, respectively. Both pension coverage and Social Security discourage saving, although Social Security tends to have a more significant impact. To estimate the magnitude of the effect of pension income on saving, the pension dummies have to be multiplied by the hypothetical saving propensities $(D - R)/LE$ which amounted to .38 in 1966 and .47 in 1969. Applying these propensities to the coefficients in the fully specified equation (3) of table 3 reveals a negative impact on saving of $671 (.38 × $1,767) associated with coverage by a private pension plan and $1,420 (.38 × $3,737) for Social Security. For 1969–71 the comparable magnitudes were $2,431 for pension coverage and $1,815 for Social Security.

The magnitude of the Social Security effect in equations (3) and (6) are consistent with the numbers derived from a very simple example. Consider the average individual in 1966 with annual income of $8,350 and assets of $19,500 who expects a Social Security benefit equivalent to the existing maximum. His lifetime wealth can be estimated as follows:

Expected earned income ($8,350 × 13.5 years) =	$112,725
Expected Social Security benefits ($3,420 × 8.5 years) =	29,070
Assets in 1966 =	19,500
Expected lifetime wealth =	$161,295

According to the life-cycle hypothesis, an individual would consume $1/LE$ or $161,295/22 = $7,332 of this expected lifetime wealth each year.[13]

---

[11] All equations were estimated both with and without intercepts. Generally, the coefficient of the intercept term was not significantly different from zero. Furthermore, the inclusion of an intercept term had no effect on the coefficients of the other variables.

[12] On the basis of the relative income hypothesis, since blacks as a group are poorer than whites they would tend to save more than whites at any given income level (Duesenberry 1967, p. 50).

[13] This calculation implies saving out of current income of $1,018 in 1966–69 and $1,629 in 1969–71. Actual saving for the sample amounted to $1,904 in 1966–69 and $2,275 in the later period. The discrepancy between the two figures can be explained by the fact that in the empirical work saving was defined as the change in the value of total assets—including capital gains.

TABLE 3

LIFE-CYCLE SAVINGS MODEL WITH PENSION DUMMIES. 1966–69, 1969–71

Model 1

$$s_t = \alpha_1 \frac{(D - R - 1)}{LE} Y_t - \alpha_2 \frac{(D - R)}{LE} PP - \alpha_3 \frac{(D - R)}{LE} SS - \alpha_4 \frac{A_{t-1}}{LE}$$

$$= \alpha_1 INC - \alpha_2 COVPP - \alpha_3 COVSS - \alpha_4 ASS$$

EQUATION NUMBER	INDEPENDENT VARIABLES								
	INC	*COVPP*	*COVSS*	*ASS*	*COLL*	*INHER*	*RACE*	*SEMP*	R^2
	1966–69								
(1) ..	1.08	−2,392	−1,950	−0.83				.	.029
	(8.6)	(1.8)	(1.9)	(6.2)					
(2) .	0.99	−3,108	−2,769	−0.94	2,217	1,784	−746	..	.035
	(7.7)	(2.3)	(2.4)	(7.0)	(2.4)	(3.6)	(1.4)		
(3) . . .	1.03	−1,767	−3,737	−1.17	2,062	1,179	−754	4,542	.046
	(8.1)	(1.3)	(3.2)	(8.3)	(2.3)	(2.3)	(1.4)	(5.6)	
	1969–71								
(4)	1.06	−5,119	−3,124	−0.87				..	.037
	(10.2)	(3.3)	(2.6)	(7.5)					
(5)	0.92	−5,547	−3,730	−0.97	1,857	2,408	−939	.	.043
	(8.3)	(3.5)	(3.0)	(8.2)	(1.7)	(4.3)	(1.5)		
(6)	0.92	−5,172	−3,861	−1.02	1,841	2,169	−953	1,804	.044
	(8.4)	(3.3)	(3.1)	(8.4)	(1.7)	(3.7)	(1.5)	(1.8)	

NOTE.—Numbers in parentheses are *t*-statistics. The first four variables are defined as follows

$$INC = \frac{(D - R - 1)}{LE} Y_t$$

$$COVPP = \frac{(D - R)}{LE} PP$$

$$COVSS = \frac{(D - R)}{LE} SS$$

$$ASS = \frac{A_{t-1}}{LE}.$$

In the absence of Social Security, the individual's expected lifetime wealth would have amounted to \$132,225 and his consumption would have been \$132,225/22 = \$6,010 The increase in annual consumption and consequent reduction in annual saving due to the guarantee of 8.5 years of Social Security benefits is equal to \$7,332 − \$6,010, or \$1,322 This reduction is very close to the \$1,420 estimated in equation (3) Furthermore, if a similar exercise is carried out for the 1969–71 period the calculated reduction in saving due to future Social Security benefits is \$1,847[14]—remarkably close to the \$1,815 estimate from equation (6).

The magnitude of the effect of private pension coverage for 1966–69 of −\$671 seems quite reasonable The average benefit from private pension plan amounted to about \$1,550 during this period, which would imply a reduction in saving of $[(D - R)/LE]$ \$1,550, or \$592 The 1969–71 results, however, are out of line with the other estimates The average private pension benefit in 1971 was \$1,684 which would imply an annual reduction in saving of \$789—one-third of the amount indicated in equation (6)

Most of the socioeconomic variables entered the equation with predicted signs and were quite significant The exception was the race variable which was negative and not significant The equations are shown both with and without the self-employed variable to emphasize how seriously its omission biases the pension coefficient[15]

Model 2.—On the assumption that pension income is proportional to earnings, the pension dummies were multiplied by actual income to provide a proxy for the level of expected pension benefits The results of these regressions are presented in table 4 Multiplying by income seems to improve the performance of the socioeconomic as well as the pension dummies Particularly significant is the impact on saving of the desire to leave an inheritance when this dummy variable is scaled by income

---

[14] The 1969–71 calculation is based on an average age in 1969 of 54 3, life expectancy of 19 9 more years, and an expected retirement age of 64 8 Total lifetime wealth amounted to 10 5 years of earnings at \$9,800, 9 5 years of Social Security benefits at \$3,888, and assets of \$23,580 for a total of \$163,416 Consumption with Social Security income would amount to \$163,416/20 or \$8,171 per year In the absence of Social Security, only \$126,480/20 or \$6,324 would be consumed annually The reduction in saving due to Social Security amounts to \$1,847 (\$8,171 − \$6,324) per year

[15] Our original model was of the form $s = b_1\ INC + b_2\ COVPP + b_3\ COVSS + b_4\ ASS$, whereas we should have been estimating $s = \beta_1\ INC + \beta_2\ COVPP + \beta_3\ COVSS + \beta_4\ ASS + \beta_5\ SEMP + \epsilon$ Using the conventional notation, the bias on the pension variable caused by the omission of the self-employment variable can be estimated as follows $E(b_2) = \beta_2 + P_{52}\beta_5$, where P_{52} is the slope coefficient in the regression of the excluded self-employment variable, *SEMP*, on the included independent variable, *COVPP*. Since self-employment and pension coverage are strongly negatively correlated and β_5 is large and positive, the coefficient of pension coverage is significantly biased downward and the negative effect of pension coverage on saving is seriously overstated As expected, therefore, the inclusion of the *SEMP* variable in the equation reduces the bias on the *COVPP* variable and significantly reduces the coefficient of the pension variable.

TABLE 4

LIFE-CYCLE SAVINGS MODEL WITH DUMMIES MULTIPLIED BY INCOME. 1966–69, 1969–71

Model 2

$$s_t = \beta_1 \frac{(D - R - 1)}{LE} Y_t - \beta_2 \frac{(D - R)}{LE} PP \; Y_t - \beta_3 \frac{(D - R)}{LE} SS \; Y_t - \beta_4 \frac{A_{t-1}}{LE}$$

$$= \beta_1 INC - \beta_2 COVPPY - \beta_3 COVSSY - \beta_3 ASS$$

EQUATION NUMBER	INDEPENDENT VARIABLES								R^2
	INC	*COVPPY*	*COVSSY*	*ASS*	*COLL Y*	*INHER · Y*	*RACE · Y*	*SEMP · Y*	
	1966–69								
(1) .	1.25 (4.9)	−0.31 (2.2)	−0.22 (1.0)	−0.78 (5.9)		.	..	...	.026
(2)	1.11 (4.4)	−0.43 (3.1)	−0.51 (2.4)	−1.12 (8.3)	0.20 (3.5)	0.42 (8.7)	−0.16 (2.5)	. .	.061
(3) .. .	1.03 (4.1)	−0.23 (1.6)	−0.59 (2.8)	−1.36 (9.6)	0.17 (3.0)	0.39 (7.9)	−0.14 (2.2)	0.42 (5.4)	.071
	Modified Social Security Variable								
(4) ...	0.92 (4.7)	−0.27 (1.9)	−0.84 (3.6)	−1.40 (9.9)	0.12 (1.9)	0.40 (8.1)	−0.12 (1.8)	0.39 (5.0)	.073

				Modified Social Security and Permanent Income					
(5)	1.19 (6.6)	−0.19 (1.4)	−1.25 (5.6)	−1.74 (12.0)	0.05 (.9)	0.35 (7.7)	−0.11 (1.8)	0.55 (7.4)	.098
				1969–71					
(6)	1.36 (6.3)	−0.43 (2.8)	−0.55 (2.5)	−0.88 (7.4)					.090
(7)	0.57 (2.3)	−0.43 (2.7)	−0.08 (0.4)	−1.11 (9.3)	0.26 (4.3)	0.34 (6.6)	−0.12 (1.9)		.056
(8)	0.52 (2.2)	−0.37 (2.3)	−0.08 (0.4)	−1.18 (9.4)	0.26 (4.3)	0.33 (6.4)	−0.12 (1.8)	0.16 (1.7)	.058
				Modified Social Security Variable					
(9)	0.90 (4.7)	−0.46 (2.9)	−0.83 (3.3)	−1.19 (9.6)	0.20 (3.2)	0.31 (6.2)	−0.11 (1.7)	0.13 (1.4)	.062
				Modified Social Security and Permanent Income					
(10)	1.28 (6.0)	−0.68 (3.9)	−1.12 (4.3)	−1.30 (10.2)	0.18 (2.6)	0.37 (6.8)	−0.11 (1.5)	0.08 (0.8)	.069

NOTE.—Numbers in parentheses are *t*-statistics

In the completely specified third equation for 1966–69, Social Security benefits once again have a larger and more significant impact on saving behavior than benefits from private pension plans. This seems quite reasonable since Social Security benefits are guaranteed, while private pension benefits are dependent on complicated vesting procedures. If some adjustment could be introduced for those persons having fully vested benefits, the significance of the impact of vested benefits on saving would probably approach that of Social Security.

To calculate the quantitative impact of Social Security and private pension benefits on saving in equation (3), the coefficients of the pension coverage variables must be multiplied by $[(D - R)/LE] \cdot Y_t$ or .38 × $8,350. This calculation reveals a negative impact on saving of $730 for private pension coverage and $1,872 for Social Security. These result are very similar to those in table 3 and are quite reasonable

The 1969–71 results are remarkably consistent with those for the earlie period with the exception of the coefficient of Social Security, which is n longer significantly different from zero once the socioeconomic variable are included in the equation In an attempt to construct a closer prox for Social Security income, the coverage dummy was multiplied by discontinuous income variable. For 1966, the Social Security dumm was multiplied by the individual's actual income if income was le than $6,600 and by $6,600 for all larger incomes. For 1969–71, th comparable figure was $7,800[16] When this improved proxy for Soci Security income is included in equations (4) and (9) of table 4 tl coefficients for the two periods are both close to − 1 and are significantl different from zero

Equations (4) and (9) indicate an annual reduction in saving of $8 for pension coverage and $1,684 for Social Security in 1966–69 ar $2,119 for pension coverage and $2,477 for Social Security in 1969–7 These figures are slightly higher than expected, but within a complete reasonable range. The exception is the large increase in the impact private pensions from 1966–69 to 1969–71

A final modification to the model was introduced in equations (5) a (10) in table 4 where current income was replaced by a simple meas of permanent income The inclusion of permanent income increases t explained variance, while the coefficients of both the Social Secur and pension income become more significant. As expected, the coeffici of the Social Security income variable is larger and more signific than that of private pension plans.

---

[16] No individual retiring in 1966–69 could have actually had a retirement ber based on the full $6,600 The highest average covered earnings for someone retirin 1966 who had always paid the maximum tax would have been $4,800. Likewise, no retiring in 1969–71 would have had average covered earnings of $7,800, as the maxim figure would have been $5,357 for any individual retiring in 1969. The $6,600 and $7 figures were used to differentiate between those who would be entitled to the maxim benefit in the periods 1966–69 and 1969–71 and those who would not.

TABLE 5

Simple Regressions 1966–69, 1969–71

Model 3

$$S_t = \gamma_1 + \gamma_2 Y_t - \gamma_3 PP - \gamma_4 SS - \gamma_5 A_{t-1}$$

Equation Number	Independent Variables									R^2
	k	*Y*	*PP*	*SS*	A_{t-1}	*COLL*	*INHER*	*RACE*	*SEMP*	
					1966–69					
(1)	−1,124	0 73	−2,361	−1,050	−0 046					062
	(1 0)	(13 1)	(3 9)	(0 9)	(7 1)					
(2)	−1,063	0 72	−2,425	−1,168	−0 047	225	625	−582		063
	(0 9)	(12 2)	(4 0)	(1 0)	(7 2)	(0 2)	(1 1)	(1 0)		
(3)	−2,423	0 76	−1,594	−1,082	−0 057	312	516	−177	4,312	073
	(1 9)	(12 7)	(2 5)	(1 0)	(8 4)	(0 3)	(0 9)	(0 3)	(5 2)	
					1969–71					
(4)	1,133	0 66	−1,710	−3,397	−0 053					054
	(0 8)	(11 0)	(2 3)	(2 6)	(8 3)					
(5)	1,001	0 65	−1,797	−3,560	−0 054	−404	1,128	−769		055
	(0 7)	(10 2)	(2 4)	(2 7)	(8 4)	(0 4)	(1 6)	(1 1)		
(6)	358	0 66	−1,462	−3,474	−0 058	−451	1,065	−589	2,197	057
	(0 2)	(10 3)	(1 9)	(2 6)	(8 7)	(0 4)	(1 5)	(0 8)	(2 1)	

Note —Numbers in parentheses are *t*-statistics

Model 3.—Finally, saving relationships were estimated for both the 1966–69 and 1969–71 periods based on simple linear saving functions which ignore the complications of age, life expectancy, and expected retirement plans The results of these regressions, which are summarized in table 5, are encouraging only insofar as they confirm that private pensions as well as Social Security coverage tend to discourage saving

Several characteristics of these results indicate that they are definitely inferior to those obtained from the more complete model First, the saving propensities are extraordinarily high and decrease with the age of the sample, whereas the life-cycle model would predict a higher propensity to save for the later period Second, the negative impact of pensions is larger and more significant in 1966–69 while the reverse is true for 1969–71 Furthermore, the negative impact of private pensions decreases with time and the age of the sample, which again is contrary to the theory Finally, the difference in the impact of Social Security between the two periods is very large and difficult to explain In short, the more complete formulation of the model presented in tables 3 and 4 provides a considerably more sensible picture of saving behavior.

IV. Conclusion

The evidence from the Labor Department sample of men aged 45–59 is definitely consistent with the hypothesis that pension coverage discourages saving in other forms When classified by age, income, and net worth, except for the lowest income class, individuals not covered by pensions had greater assets than those expecting benefits Furthermore, in all regression equations, coverage by pension plans had a negative coefficient These new results directly contradict the findings of earlier studies by Cagan (1965) and Katona (1965) that persons covered by private pensions saved more than those not covered

This study, however, differs from earlier ones in important respects which may explain why previous efforts have failed to pick up the effect found here We have focused solely on the saving behavior of heads of households in the preretirement years For these people, providing for retirement is the primary saving motive, and, therefore, these should be the people whose saving behavior is most sensitive to pension coverage Furthermore, the data used in this study are from a later period when pension coverage was more important Pension plans have tripled in size since 1960 when the earlier studies were undertaken and, therefore, individuals may simply have become more knowledgeable and sophisticated in their saving decisions

Since the sample under consideration is representative of the nation, it is reasonable to extrapolate the behavior of this group to the entire country and derive macroeconomic implications In 1973, there were

27.5 million workers covered under private retirement plans (Kolodrubetz 1975, p. 16, table 1). If each of these workers expected to retire at age 65 with an annual pension of $1,830 (the average benefit in 1973, see table 1) and if the propensity to save out of these expected benefits equaled the ratio of expected retirement span to life expectancy for each age group, then the reduction in annual aggregate saving in 1973 due to future private pension benefits would be about $13 billion. The reduction in saving would be less if younger groups are very insensitive to pension coverage. Total contributions to private pensions amounted to $21.1 billion in 1973, highlighting the fact that the substitution of pension saving for other saving is imperfect probably because of lack of vesting and other uncertainties. On the basis of these rough calculations, private pension plans appear to increase aggregate saving on net and contribute to capital accumulation.

The implications of the negative impact of Social Security are considerably different. Unlike private pensions, Social Security is not funded and therefore a reduction in individual saving in anticipation of future Social Security benefits results in a net decline in aggregate saving. These results indicate that Social Security does have a significant negative impact on saving which confirms the findings of two earlier studies of aggregate saving and Social Security by Munnell (1974) and Feldstein (1974).

References

Ando, Albert, and Modigliani, Franco. "The 'Life-Cycle' Hypothesis of Saving: Aggregate Implications and Tests." *A.E.R.* 53 (March 1963): 55–84.

Cagan, Phillip. *The Effect of Pension Plans on Aggregate Saving: Evidence from a Sample Survey.* National Bureau of Economic Research, Occasional Paper no. 95. New York: Columbia Univ. Press, 1965.

Duesenberry, James. *Income, Saving, and the Theory of Consumer Behavior.* Cambridge: Harvard Univ. Press, 1967.

Feldstein, Martin S. "Social Security, Induced Retirement, and Aggregate Capital Accumulation." *J.P.E.* 82, no. 4 (September/October 1974): 905–26.

Gold, Ronald B. "Tax Deductions for Individual Retirement Savings." *Nat. Tax J.* 25 (December 1972): 585–93.

Katona, George. *Private Pensions and Individual Savings.* Monograph no. 40, Survey Research Center, Institute for Social Research. Ann Arbor: Univ. Michigan, 1965.

Kolodrubetz, Walter W. "Employee Benefit Plans, 1973." *Social Security Bulletin* 38 (May 1975): 22–29.

Modigliani, Franco, and Brumberg, Richard. "Utility Analysis and the Consumption Function: An Interpretation of Cross Section Data." In *Post Keynesian Economics,* edited by Kenneth Kurihara. New Brunswick, N.J.: Rutgers Univ. Press, 1954.

Munnell, Alicia H. *The Effect of Social Security on Personal Savings.* Cambridge, Mass.: Ballinger, 1974.

Slavick, Fred. *Compulsory and Flexible Retirement in the American Economy*. New York State School of Industrial and Labor Relations, Cornell University. Ithaca: Cornell Univ. Press, 1966.
U.S. Department of Health, Education, and Welfare, Public Health Service. *Vital Statistics of the United States 1970.* Volume 2. *Mortality*, Part A. Rockville, Md.: Nat Center for Health Statis., 1974
U.S. Department of Labor. *The Older American Worker. Age Discrimination in Employment* Washington Government Printing Office, 1965.
U S Department of Labor, Manpower Administration. *The Pre-Retirement Years* Vol. 1 Manpower Research Monograph no 15 Washington: Government Printing Office, 1970
U S Department of Labor, Manpower Administration *The Pre-Retirement Years* Vol 3. Manpower Research Monograph no 15 Washington Government Printing Office, 1973.

The Economics of Information and Retail Gasoline Price Behavior: An Empirical Analysis

---

Howard P. Marvel

Ohio State University

This paper provides empirical results which indicate that consumer responses to imperfect, costly information have an important impact on price behavior in the retail gasoline market Two aspects of price behavior are investigated the market price dispersion at a point in time and the variability of price over time. Through use of a multiple-regression cross-city analysis of price behavior, both of these characteristics are shown to depend on a set of proxy variables representing the benefits and costs to consumers of acquiring information In addition, prices vary more within a given city at stations catering to relatively well informed customers

The burgeoning literature in the economics of information has yielded a large number of testable hypotheses that relate prices and information, yet few empirical results are available to indicate that information accounts for more than "variations in the numbers we observe at the fifth or sixth decimal point "[1] This paper attempts to fill this gap in our knowledge by demonstrating that in one important retail market, that for gasoline, information has a significant impact on two aspects of price behavior the amount of offer price dispersion one observes among stations within a single market and the extent of price variability over time Section I contains results of a regression analysis of the relation between offer price dispersion and a set of variables designed to serve as proxy

---

I am grateful to John P Gould, George J Stigler, Dan Wisecarver, an anonymous referee, and especially Lester G Telser for helpful comments Any errors are my responsibility

[1] Rothschild opens his excellent survey article (1973) with this challenge from a colleague to show that information's impact is important He attempts to answer the challenge by summarizing the theoretical applications of the concept of imperfect, costly information This paper is intended to complement his discussion by answering that challenge empirically

[*Journal of Political Economy*, 1976, vol 84, no 5]
© 1976 by The University of Chicago All rights reserved

measures for the amount of information consumers choose to acquire. The estimated coefficients, while not always statistically significant, consistently exhibit the signs predicted by information theory Given certain deficiencies of the data and special characteristics of the retail gasoline market, these estimates are likely to provide a lower bound to the impact of information on retail price dispersion

The results given in Section 2 indicate that there is a strong relation between information and price variability over time It is argued that markets with well-informed customers will experience relatively large amounts of temporal price variability, both because the input of uncertainty to such markets is large and because shocks introduced to such markets are more widely transmitted than in markets with less well informed customers The empirical results presented in support of the hypothesis are clear, consistent, and significant

1. Sources of Offer Price Dispersion

The derivation of empirically testable hypotheses concerning the dispersion of offer prices from the theory of information is a two-step process Since observations on the size of consumer information holdings are not generally available, it is necessary to represent such stocks by proxy This can be done by investigating the costs and returns to consumers of acquiring information in order to obtain a set of observable variables affecting a consumer's optimal stock of information The second step is to develop the linkage between consumer information and observed dispersion Stigler (1968, p 178) argues that "greater amounts of search will lead to a smaller dispersion of observed selling prices by reducing the number of purchasers who will pay high prices " Unfortunately, as Rothschild (1973) has stressed, imperfect consumer information is not a sufficient condition for price dispersion A complete model of price dispersion must include a specification of how firms' price-setting behavior responds to changes in the amount of knowledge consumers have In other words, one needs a theory of the equilibrium price distribution to be able to assess the impact of information on price dispersion in markets for different goods Such theories exist, but they are in their infancy and provide little guidance for empirical testing [2]

The problem of specifying the determinants of the equilibrium distribution can be avoided here because I wish to assess the partial effect of differences among consumers on price dispersion across markets for a single good, gasoline Since gasoline is marketed in a similar fashion throughout the United States, most information-based differences in dispersion will be due to differences in customers, not firms To allow for variations in firm response arising from differences in market structure, the Herfindahl index of seller concentration appears as an independent variable in the regression equation Beyond this adjustment, it is assumed that the information dispersion linkage is the same in all cities

---

[2] Many of these models are surveyed in Rothschild (1973) One of the most interesting models not covered by Rothschild is in Telser (1974), see also Marvel (1975)

While the relation between information and dispersion is not well understood, the determinants of the amount of information consumers choose to obtain have been studied extensively Since the results are familiar, one need only state them and then specify an appropriate set of proxy variables [3] The proxies used here are of two types those affecting the gross gain from search and those determining search cost The gross gain from search is the expected reduction in price paid times the number of units the individual expects to purchase at the lower price which search makes possible Telser (1973) has demonstrated that shifting the entire distribution of offer prices has no effect on the expected number of searches carried out, given that the number of units purchased is fixed This suggests that the gross gain from search will be largest in those cities where per capita gasoline consumption is relatively high,[4] so that observed price dispersion in such cities will be small

The gross gain from a given amount of search activity also depends on the time interval over which the information obtained remains valid If a consumer expects the savings generated by search to apply to a large number of successive purchases, the gross gain to him from search will be greater than if the price savings apply only to the current purchase Stigler uses this argument to explain why tourists can be exploited· it does not pay for a tourist to search out low prices, since the gain from obtaining a low price quotation applies to only one purchase Similarly, in the gasoline market we expect the price behavior of high-volume stations located on major intercity routes to be quite different from that observed at stations on intracity arteries frequented by the same buyers over time Even among stations of the same type, we predict that consumers will choose to hold smaller stocks of information the lower the correlation of firms' offer price locations within the price distribution in successive time periods [5] It turns out to be quite important in the empirical investigation to deal with the impact of the depreciation rate of price information on the amount of information consumers possess

The cost of search is composed both of the cost of goods employed in search—gasoline, automotive depreciation, and the like—and of the cost of time spent in searching. This latter cost is probably a principal component of search cost, and since the price of time (the wage rate for employed persons) is closely related to income, it seems that income will serve as a good proxy for search cost. One problem with this measure is that the link between income and search is weakened by differences among consumers in their efficiency of search. For example, if we compare two individuals, one who commutes by a similar route every workday and another whose driving pattern is irregular, the commuter will be able to use his information for a larger number of purchases and in that sense will be more efficient in obtaining valuable information. Similarly, we expect income and education to be correlated, so that if education raises the efficiency of search, an education variable should be included in a multiple-regression study of price dispersion. We would expect a rise in education level, holding constant the prices of time and goods used in search, to lead to increased search efficiency and so to result in an increase in consumer information stocks Including an education variable will also serve to reduce any efficiency effects caught by the income variable and so will improve income's usefulness as a proxy for the cost of search

---

[3] See Marvel (1975) for a derivation of the results

[4] Measures of units consumed are likely to be superior to dollar sales measures for goods like gasoline which have inelastic demand in the relevant price range.

[5] Consumers may actually carry out more searches by searching at shorter intervals if the depreciation rate increases Nonetheless, their reservation price will rise. The distinction here is between investment flow (number of searches) and capital stock (as indicated by the reservation price) The latter is relevant for my purpose

In summary, for a given initial distribution of offer prices, the range of acceptable prices will be smaller (1) the larger the volume of purchases, (2) the greater the correlation of successive prices in the price distribution, (3) the lower the buyer's income level, and (4) the higher the buyer's educational attainment

I now turn to estimates of the impact of consumer information stocks on observed retail gasoline offer price dispersion These estimates are of equations relating city-by-city price dispersion to the proxy variables representing consumer information outlined above, to the Herfindahl index of seller concentration, and to a measure of tax rate differences across the urban area in question. As noted above, the Herfindahl index appears as a shift variable to allow for differences in retail market structure across cities, it summarizes the relative size of sellers in the market as measured by the various brands' market shares of sales.[6]

The empirical results are based on price data made available by the U S Bureau of Labor Statistics (BLS). These data are maxima and minima of samples of gasoline pump prices collected for use in the preparation of the Consumer Price Index (CPI). Prices of both regular and premium

---

[6] Let s_j denote the market share of unit sales of firm j in a given market If n firms supply that market, the Herfindahl index, H, is defined as

$$H = \sum_{j=1}^{n} s_j^2 .$$

The share data are derived from state motor fuel tax records by the Lundberg Survey, Los Angeles, and are reported in the 1970 mid-May factbook issue of *National Petroleum News* (1971). Shares are therefore of statewide sales. For urban areas overlapping state boundaries, the H index is a weighted average of the state indices corresponding to the SMSA, the weights being the share of total SMSA gallonage sold in that state.

grade gasoline are obtained from stations in the BLS samples, subject to availability Samples are collected monthly in 10 Standard Metropolitan Statistical Areas (SMSAs) and quarterly in 13 others. The BLS was willing to provide only the extrema from both the regular and the premium gasoline samples along with the corresponding sample sizes The data therefore consist of four time series (two per grade) of gasoline prices for each of 23 cities The data are available for the period subsequent to the last major revision of the CPI in 1964 through the middle of 1971.

The cross-city analysis of price dispersion has been carried out by compressing the time series of sample ranges of prices for each city and for each gasoline grade into a single measure of price dispersion There are several pitfalls in the design of such a measure The BLS attempts to obtain prices from the same set of stations in successive time periods. Since the location of a particular station's price in the offer price distribution remains relatively stable over time, successive BLS samples are far from the independent drawings of offer prices which would be ideal for the purpose of estimating "normal" price dispersion If an outlier—a station charging an especially high or low price—is included in one sample, it will likely continue as an outlier in succeeding samples. The problem is compounded by the fact that dispersion must be estimated by the range, a statistic subject to substantial sampling error One cannot expect to reduce sampling error very much by averaging sample ranges in some fashion The net effect of these considerations is that any measure of dispersion based on the available data will be quite imprecise, suggesting that the estimates of information's impact on dispersion are likely to constitute a lower bound on the strength of the relation

A further problem involving use of these ranges arises because of variation in the size of samples obtained by the BLS The sample range is quite sensitive to sample size, and since sample sizes differ both within the city time series and across cities, some allowance must be made for these differences Unfortunately, since gasoline price distributions are typically skewed,[7] the usual procedure of adjusting for sample size with Patniak's approximation is not appropriate [8] To minimize bias due to sample size fluctuation, only ranges from samples of the most common size—12 observations—were used Thus, dispersion is measured by the average deflated range for samples of size 12 [9]

---

[7] Examples of gasoline offer price distributions are reported in Cassady (1954), Livingston and Levit (1959), and Allvine and Patterson (1972)

[8] See, e g , Duncan (1965, p 131) The range of samples from any distribution is approximately χ^2, but Patinak's conversion factors are based on normality of the underlying population

[9] Let $S(t)$ give the sample size obtained for a given time series in month t Define $A = \{a\ S(a) = 12\}$ The measure of dispersion is just

$$r = \frac{1}{n(A)} \sum_{a \in A} \frac{P_{Ha} - P_{La}}{x_a},$$

where $P_{Ha} - P_{La}$ is the range at time a, $n(A)$ gives the number of elements in A, and x_a

A look at these ranges indicates substantial differences in price dispersion among the cities in the sample—differences larger than could reasonably be attributed to differences in retailing methods. The average premium range for Saint Louis is 9¢ and that for Cleveland is 1.25¢. The mean range is 4 8¢ for premium and 4.5¢ for regular, and the respective standard deviations are 1 7¢ and 1 5¢

Table 1 contains estimates of a regression equation relating these ranges to information proxies. The estimates are of the following form:

$$\ln r_i = \beta_0 + \beta_1 \ln b_i + \beta_2 \ln y_i + \beta_3 \ln s_i + \beta_4 \ln v_i + \beta_5 \ln m_i + \beta_6 \ln t_i + \delta_i, \qquad (1)$$

where $i = 1, \ldots, 22$ is an index indicating city i and r_i = average range, b_i = gasoline consumption per car, y_i = median family income, s_i = median years schooling,[10] v_i = a measure of gasoline price variability over time, m_i = the Herfindahl index of market structure, t_i = a correction factor allowing for differences in tax rates across the SMSA,[11] and δ_i = disturbance term

The first four independent variables are the proxy variables for the amount of information consumers choose to obtain. Since an increase in per car purchases of gasoline will result in greater benefits from a given amount of search, cities with high gasoline consumption rates will be expected to have well-informed gasoline consumers, so β_1 should be negative.[12] Income is included as a measure of the cost of search and

---

is the corresponding BLS city index of consumer prices The use of this measure meant that Honolulu had to be excluded (no samples of size 12) and that data from New York, Chicago, and Los Angeles were derived from an earlier period (1964–66) than those from the remaining 19 SMSAs (1966–71).

[10] The gasoline consumption figure is the sum of estimated gallons sold for SMSA constituent counties (For Boston the Metropolitan State Economic Area, MSEA, was used in place of the SMSA) Data are from the 1967 Census of Business (U S Bureau of the Census 1971*a*, 1971*b*, 1971*c*) The number of automobiles is obtained from the 1970 Census of Population and Housing (U S Bureau of the Census 1972*a*, 1972*b*, 1972*c*) This same source yielded the income and schooling figures

[11] Tax differences arise when SMSA boundaries overlap portions of states with different gasoline tax rates or when gasoline is subject to percentage rate sales taxation Let d_i be the difference between the highest and lowest effective taxes applied to a given grade of gasoline in SMSA i Then the tax rate adjustment t_i is defined to be $1 + d_i$ That is, ranges not subject to tax differences are unaffected while those for SMSAs exhibiting tax differences are scaled down The tax rate coefficient is expected to be positive. Sales tax policy of the states in the sample was obtained from the Commerce Clearing House *State Tax Guide* (1967) and the Census Bureau's *State Tax Collections* (1971*c*) Contacts with the revenue authorities of the states represented in the sample provided the rates in force over the period Gasoline excise rates were obtained from the American Petroleum Institute, *Petroleum Industry Factbook* (1971)

[12] The referee for this paper has suggested that the observed significant relation between gasoline volume and price dispersion may simply be measuring a demand for gasoline relation Since gasoline price distributions typically exhibit negative skewness, the price which governs most purchase decisions is the high price from the BLS samples The correlations between high prices and price dispersion are generally insignificant and always negative, suggesting that the volume-dispersion relationship is not a demand relation.

TABLE 1

PRICE DISPERSION REGRESSIONS*
(22 CITIES)†

DEPENDENT VARIABLE PRICE SERIES	COEFFICIENTS (*t*-STATISTICS IN PARENTHESES)						
	Constant	ln Gallons per Car	ln Median Family Income	ln Median Years Schooling	ln Herfindahl Index	ln Var (Δ ln p)‡	Tax Rate Adjustment
Premium	−4 857 (−0 78)	−2 198 (−1 79)	0 846 (0 89)	−2 550 (−0 75)	−0 393 (−1 10)	0 295 (2 46)	19 70 (2 35)
Regular	−6 052 (−1 23)	−2 880 (−2 98)	0 538 (0 73)	−0 870 (−0 34)	−0 190 (−0 67)	0 308 (3 27)	31 70 (4 38)

---

* Two-stage least-squares estimates of text equation (1)

† Honolulu excluded—no samples of size 12

‡ Low-price series, same grade as dependent variable

schooling as a measure of search efficiency; consequently, we expect to observe $\beta_2 > 0$ and $\beta_3 < 0$.

The remaining information proxy, gasoline price variability, poses certain special difficulties. First, it is not precisely the variable that is needed. Some measure of the depreciation rate of information stocks must be included in the regression equation if the independent variables are to serve as proxies for information.[13] Ideally this variable should be a measure of the autocorrelation of station price locations within the offer price distribution over succeeding samples The lower this autocorrelation, the greater the deterioration of consumer information holdings over time and, hence, the smaller the equilibrium stock of information. This rate of depreciation may well be correlated with the amount that the entire distribution shifts Given that firms are uncertain as to the proper price to charge, one expects their uncertainty to be greater and their need to change their own prices to be more pressing the larger the fluctuations in the whole price distribution Unfortunately, the relation between price variability over time, on the one hand, and the stability of locations of particular stations in the price distributions, on the other, will be far from perfect For example, a change in the tax rate in a particular locale, being known with certainty to all firms, is likely to result in an upward shift of the entire distribution such that the relative locations of all of the stations in that market are maintained. An exogenous increase in demand which causes prices to rise is much more likely to have a differential impact across firms and so lead to changes in their relative locations in the price distribution. In general, one can expect that changes in demand conditions will lead to greater depreciation of price information than will changes in supply conditions, since the latter, being determined on the national level, are more likely to have a uniform impact on all firms in the market For this reason, the variability of low prices over time has been used as an independent variable[14]

A second difficulty involving the use of price variability over time as an independent variable is that it is not exogenous A firm's policy regarding frequency and amplitude of its price movements will depend on how quickly customers respond to such price changes—and these response patterns will be governed by consumer knowledge Accordingly, equation (1) was estimated using two-stage least squares[15]

---

[13] In the absence of a price information depreciation variable, the remaining coefficients will be biased for reasons indicated below

[14] I will show (see table 6) that high prices are much more closely related to national market conditions than are low prices

[15] The system consists of eqq (1) and (2) Since the system is recursive, eq. (2) can be estimated with ordinary least squares (OLS) Unfortunately, since information is in both cases measured by proxy, it is unlikely that the covariance of the disturbance terms of the two equations will be zero, so OLS is inappropriate for (1) (see Fisher [1970, pp 13–15]) Another problem with estimating eq. (1) arises because the excluded variable, the vari-

I have pointed out that the remaining variable, the Herfindahl index, appears to ensure that differences across markets in price dispersion reflect differences in the characteristics of gasoline consumers in those markets If firms as well as customers are imperfectly informed so that no firm knows the "best" price for it to charge, potential disagreements about prevailing market conditions, and hence about prices, might be expected to be greater the larger the number of independent judgments made concerning such conditions The number and importance of independent judgments will be inversely related to the concentration of supply in a few hands;[16] hence, we expect the coefficient of the market structure variable, β_5, to be negative This argument does not involve the state of competition in the market

Although the significance levels of some of the coefficients in table 1 are low, the signs are in every case in accordance with expectations An increase in units purchased or a decrease in search cost leads to a decrease in price dispersion, the greater the price variability over time, the larger the normal observed dispersion, and the more concentrated the market (the larger the Herfindahl index), the smaller the deserved dispersion of prices The weakest variables are the proxies for search cost—income and schooling—which should not be surprising, since the correspondence between these variables and search cost is expected to be weaker than that between the benefits from search and the proxy variable representing benefits Considering the deficiencies of both the price data and the proxy variables, these results provide empirical support for the proposition that information affects price dispersion

2. Sources of Gasoline Price Variability over Time

The results in Section 1 indicate that imperfect information accounts for a portion of the dispersion of retail gasoline offer prices observed at any particular point in time Perhaps a more significant aspect of price behavior is the way in which prices vary over time, and as this section will demonstrate, trader responses to imperfect information have an important

---

ance of Bureau of Employment Security (BES) annual unemployment rates, has a relatively small impact on price variability The result of the inability of this instrument to push the endogenous variable (low-price variability over time) is that the fitted values of the exogenous variables will be nearly linear combinations of the exogenous variables appearing in eq (1), thus leading to multicollinearity and inflated standard errors. The BES data are from the 1973 *Manpower Report of the President* (U S Department of Labor 1973).

[16] I expect disagreement to increase with the number of independent judgments for two reasons First, each decision maker is basing his decision on imperfect information. If we think of this information as the realization of a drawing from a fixed distribution, then the variance of such realizations will increase with sample size Second, firms will no doubt reduce their uncertainty by obtaining price quotations of competitors, but the cost of a complete set of such quotations will increase with the number of rivals.

impact on this type of price behavior as well. A consumer's decisions about how much price information to acquire and when to carry out his search depend in part on the price charged by his current supplier. To the extent that changes in the firm's offer price affect the timing of search, costs of changing prices are imposed on that firm. Moreover, the amount of information consumers obtain regarding prices at rival firms will determine the consequences of a firm's failure to respond to changes in demand or cost conditions It will be argued that firms with relatively well informed customers are most likely to vary prices both because the costs to them of changing prices are comparatively low and because the consequences of not doing so are high. This relation is tested in two ways Gasoline stations within an urban area will differ in the information holdings of the customers they attract. One test compares price variability at stations expected to serve well-informed customers with that observed at stations serving poorly informed consumers A second, separate test involves comparison of price variability in different urban areas This test relates a city's gasoline price variation to the characteristics of that city's gasoline customers which determine the amount of information they will choose to obtain The results provide strong support for the view that information is an important determinant of price variability over time

A firm's decisions as to what offer price to set and when to change its price depend on the anticipated responses of both its actual and its potential customers. Consider first the impact on a firm's current clientele of a change in that firm's offer price The firm's customers have presumably settled on that firm following a search of their options and have no reason to collect additional information unless they suspect that the offer price distribution may have changed One obvious signal that conditions have altered is provided by a change in their supplier's offer price If the new offer price serves to induce additional search, the firm can expect some customer defections as this search turns up alternative suppliers offering more attractive terms. The defections are presumably greatest for firms in the upper portions of the offer price distribution, since their customers are the most likely to discover lower prices through search Firms with customers who are, on average, relatively well informed are able to attract such customers because they offer relatively advantageous terms of sale Their defections caused by additional search would therefore be smaller, since the probability of discovering a better alternative would be low A firm in the upper tail of the offer price distribution has obtained customers who have gathered relatively small amounts of information, so that firm must weigh carefully the costs to it of actions which may induce additional consumer search.

Balancing this reason for not altering prices is the potential loss incurred if the firm holds its price stable in the face of evidence that the price it has chosen is incorrect Obviously, a price set too low causes forgone profits on current volume and leads to an undesirable increase in sales volume, while a too-high price implies customer defections But notice that the size of such penalties is a function of the flow of customers from outlet to outlet A firm serving a poorly informed clientele can, for short periods, maintain a price which is above its normal position in the offer price distribution because its customers will be slow to discover the disparity Similarly, a price which is temporarily too low will not induce a flow of well-informed customers unless that price places the firm in the lower tail of the offer price distribution The need to respond to underlying changes in market conditions is thus less urgent for those firms facing poorly informed customers Moreover, to the extent that movements of customers between firms provide information to those firms about conditions prevailing in the market generally, firms catering to comparatively well informed customers will be the first to find out about changes in conditions which require them to alter prices Firms serving poorly informed customers are the least likely to discover small market disturbances that lead some firms to change prices

The net effect of these considerations is to predict that prices will vary most at those firms catering to relatively well informed buyers If at a given point in time the location of a particular firm's price in the offer price distribution were random, there would be no reason to expect systematic differences in the type of customer served In the retail gasoline market, however, the relative locations of particular sellers' prices in the offer price distribution persist over time [17] If not all gasoline customers are fully aware of these price differentials, low-price stations will tend to have customers with relatively good knowledge of offer prices simply because additional information increases the probability of discovering a low price But even if customers have a general awareness of prevailing price differentials, there is still reason to expect customers of various stations to differ in significant ways One such difference involves the type of driving done High-volume, low-price stations typically locate along major intracity arteries and therefore are more likely to be frequented by commuters than are high-price, low-volume neighborhood stations We expect that commuters will choose to obtain more price information than will typical customers of high-price stations, both because commuters receive greater benefits from such information and because their search costs are relatively low The low search costs arise simply because stations can be canvassed along the route taken to work with only slight additional effort and delay. On the other hand, customers of neighborhood stations may have to make special trips in order to canvass additional sellers.

---

[17] Gasoline prices from 18 stations in Akron, Ohio, for 1953–55 are contained in Learned and Ellsworth (1959) The low price over the sample period occurred at one of two independent stations, while the high price was always found at one of two major stations The normal difference between the prices at particular stations and the sample mean price was quite stable The Akron data also contain persistent price differences among stations selling the same brand of gasoline

Price information confers larger benefits on commuters for several reasons If one assumes that commuters drive longer distances and hence use more gasoline than customers of neighborhood stations, they receive larger benefits from a given decline in price per unit resulting from search Not only are such savings larger, but the commuter is also in a relatively good position to take advantage of the low prices he discovers Even if he encounters a low-price station with a full tank of gas, he retraces his steps daily and will eventually be able to use his knowledge. On the other hand, a customer of a neighborhood station is not so likely to return to a low-price station encountered when that customer had a full tank The net effect of these considerations is that commuters will be aware of more prices than customers of neighborhood stations and will be better able to respond to the low prices they encounter Commuters therefore have available more useful information, so high-volume, low-price arterial stations catering to commuters are likely to exhibit relatively large amounts of price variability

The relatively large amounts of information possessed by customers of low-price stations will not only lead those stations to respond more quickly than their high-price counterparts to shifts in supply or demand conditions affecting the market as a whole, they will also increase the supply of shocks or uncertainty facing these stations alone The reason for this involves the nature of gasoline supply arrangements at the wholesale level The wholesale market for gasoline is not a tightly organized exchange but a loose collection of brokers, marketing agents, agents for refiners, and jobbers. No one price prevails for gasoline of a particular specification, bargains are available to those who seek them, and it will pay the typically low price seller to do the searching Since his demand schedule is relatively elastic, he can sell the extra gasoline he buys without a large fall in price That same elastic demand means that he can rely on the spot market for gasoline, knowing that a short supply today can be rationed using high prices without substantial long-term consequences If a high-price station tried to rely on short-term purchases rather than on the characteristic long-term supply arrangement with a major oil company, that station might find that, once it has lost customers due to a temporary supply shortfall and consequent higher price, those customers will take some time to return when its price returns to a more normal level The operator of a high-price station is in business to provide a stable supply of a well-known brand of gasoline at a reasonably stable price, thereby sparing his customers the need to carry out a considerable amount of search His policy is similar to that of any supplier holding inventories to buffer prices for his consumers The contrast with the low-price stations

—their price variability and their uncertain gasoline availability—is apparent. Each type of station is selling gasoline and information in a package appropriate to the type of customers it serves.

It has often been alleged that the retail gasoline market is characterized by collusion among major-brand gasoline suppliers. While no direct evidence on the presence of collusion is offered, the operation of a collusive agreement would affect gasoline price variability. The considerations mentioned above make low-price stations an attractive place for refiners to dump excess gasoline. We would ordinarily expect that due to price discrimination prices would be lower in the more elastic portions of the market. A monopolist would vary prices in both segments of the market to equate marginal cost with marginal revenue in each, so that we might expect the same frequency of price changes in each segment but more amplitude in the high prices. An oligopoly, however, may not be able to fine-tune prices in this manner, with its members choosing instead to dump uncommitted, unanticipated stocks of gasoline onto the highly elastic low-price market while leaving the lucrative neighborhood-station market untouched. It may be that gasoline marketers compete in the low-price segment of the market while colluding at the high-price end of the same market. It is the imperfect information of customers in the high-price end of the market that allows them to do so.

In summary, one can expect prices to be most variable at high-volume, low-price gasoline stations serving well-informed customers. This is because such stations receive the largest benefits and incur the smallest costs of changing prices and because such firms are most able to adjust to changes in the volume of gasoline they are able to obtain. This proposition is easily extended to comparisons of price variability across cities. We expect that prices will be more variable in cities where it pays gasoline customers to obtain relatively large amounts of price information. These predictions can be tested using the BLS gasoline price data introduced in Section I.

The results reported here employ two measures of price variability: the variance of first differences in prices, var (Δp), and the variance of first differences in the logs of prices, var $(\Delta \log p)$. For a given price series, p_t, $t = 1, \ldots, T$, we have

$$\text{var}\ (\Delta p) = \sum_{t=2}^{T} \frac{(\Delta p_t - \overline{\Delta p})^2}{T - 2}$$

and

$$\text{var}\ (\Delta \log p) = \frac{1}{T - 2} \sum_{t=2}^{T} (\Delta \log p_t - \overline{\Delta \log p})^2,$$

where a bar over a variable denotes its mean and Δ is the first-difference operator, $\Delta p_t = p_t - p_{t-1}$. Var (Δp) is therefore the mean square of deviations in first differences from a linear trend, $\overline{\Delta p}$, while var $(\Delta \log p)$

is the mean square of deviations of $\Delta \log p$ from a percentage trend, $\overline{\Delta \log p}$. Trends are removed because we are concerned with the way in which prices vary when moving from level to level rather than the nature of the determinants of the price levels themselves.

The two measures of price variability differ not only in terms of the trend correction employed but also in the interpretation of price variability implicit in their construction. Compared with var (Δp), the use of first differences of logs of prices gives relatively less weight to large price changes. In the context of the gasoline market, this means giving less weight to price behavior commonly termed price warfare. Deep price cuts are more likely the thinner the spot market in gasoline. Since so-called price warfare is, of course, most common at the low-price end of the market, the fewer the informed customers who move from low-price station to low-price station the thinner the low-price market and, hence, the deeper the price cuts when an external disturbance is introduced. The use of var $(\Delta \log p)$ implies that one argues that such price cutting is not indicative of the normal level of price variability.[18]

Variances of first differences in high and low prices for both regular and premium grade gasoline are presented by city in tables 2–4. Before the various measures of price variability were calculated, the BLS gasoline prices were deflated with corresponding BLS indices of consumer prices, so that the measures involve first differences in real prices. Note that the qualitative behavior of the two measures of variance, var (Δp) and var $(\Delta \log p)$, is the same, so that my interpretation of price variability is independent of the particular measure used. Given that customers of low-price stations choose to obtain more information than their counterparts at high-price stations, my hypothesis relating information and price variability receives strong empirical support. Comparing var (Δp) for high and low prices, we see that low prices are more variable than high prices in 40 out of 46 possible cases.[19] Well-informed customers therefore pay not only low prices, they also pay variable prices.[20]

---

[18] It should not be inferred from this discussion that the results presented here are particularly sensitive to the measure of price variability which is employed. They are not. Comparisons of price variability give consistent results for both variance measures.

[19] This comparison refers to the variances in tables 2 and 4. Because high and low prices in a given city will be correlated, an F-test is inappropriate for the problem of comparing these variances. In its place I have employed a significance test (due to Pitman) which allows for the correlation between the price series (see Snedecor and Cochran [1967] for details).

[20] There is a remote possibility that the results given here could reflect the fact that gasoline price distributions commonly exhibit negative skewness. Minima from random samples from such distributions will have larger variances than will maxima. But due to the BLS replicated-outlet sample procedure, samples of gasoline prices taken at different points in time are not independent drawings. If the price at a particular station is unavailable when a sample is scheduled to be collected, the actual size of that sample will be less than the scheduled sample size. Identical stations appear in successive samples if, and only if, actual and scheduled sample sizes are the same. When this happens, the price

TABLE 2

VARIANCES* OF MONTHLY FIRST DIFFERENCES IN RETAIL GASOLINE PRICES
(PITMAN'S TEST STATISTIC† IN PARENTHESES)

CITY AND GASOLINE GRADE	VAR (ΔP) High Prices	VAR (ΔP) Low Prices	VAR (Δ log P) High Prices	VAR (Δ log P) Low Prices
Chicago				
Premium	0.065	0.302	0.406	2.867
	(0.650)‡		(0.755)‡	
Regular	0.055	0.359	0.445	4.557
	(0.738)‡		(0.827)‡	
Detroit.				
Premium	0.609	0.240	5.280	2.859
	(−0.481)§		(−0.334)§	
Regular	0.574	0.255	6.276	4.003
	(−0.433)§		(−0.250)§	
Los Angeles				
Premium	0.106	1.227	0.817	12.704
	(0.846)‡		(0.883)‡	
Regular	0.050	1.335	0.459	17.428
	(0.929)‡		(0.950)‡	
New York				
Premium	0.016	0.088	0.103	0.805
	(0.727)‡		(0.803)‡	
Regular	0.015	0.117	0.115	1.353
	(0.777)‡		(0.849)‡	
Philadelphia				
Premium	0.013	0.014	0.097	0.147
	(0.049)		(0.234)‡	
Regular	0.017	0.018	0.162	0.233
	(0.036)		(0.193)	

NOTE.—Prices are deflated by BLS city price indices.
* Reported values are actual variances × 10^3.
† See n. 19 herein.
‡ Variability of low prices is significantly greater than that of high prices at the 1 percent level.
§ Variability of high prices is significantly greater than that of low prices at the 1 percent level.

Prices not only vary less at high-price stations than at low-price stations, they also contain different sorts of serial dependencies. For the high-price series, the figures for var (Δp) calculated from quarterly first differences are generally two to three times greater than the magnitude of those for monthly first differences.[21] The ratio of the quarterly figure to the monthly figures is generally largest for those series with small variances, and in most cases the ratio for high prices exceeds that for low prices. This observation is consistent with the hypothesis that high prices follow

---

changes are clear of variation, due to changes in the identity of the sampling units. The variances in tables 2–4 were recalculated using first differences between months when actual sample size equaled scheduled sample size (results available on request). These results do not compromise the conclusions drawn from tables 2–4.

[21] These statements are based on evidence for 10 cities—the five cities in table 2 and the remaining five cities for which BLS gasoline price samples are collected monthly. For the second group of cities, comparisons of variances of first differences of undeflated prices were made. This has little effect on the results.

TABLE 3

VARIANCES* OF QUARTERLY FIRST DIFFERENCES IN RETAIL GASOLINE PRICES
(CITIES FOR WHICH MONTHLY DATA ARE AVAILABLE, PITMAN'S TEST STATISTIC† IN PARENTHESES)

CITY AND GASOLINE GRADE	MEASURE OF VARIABILITY	(1)‡		(2)‡		(3)‡		AVERAGE	
		High Prices	Low Prices	High Prices	Low Prices	High Prices	Low Prices	High Prices	Low Prices
Chicago.									
Premium	Var (Δp)	0.135	0.850	0.134	0.545	0.115	0.425	0.128	0.607
		(0.733)§		(0.613)§		(0.596)§			
	Var (Δ log p)	0.831	8.385	0.822	4.922	0.707	4.107	0.787	5.805
		(0.824)§		(0.720)§		(0.726)§			
Regular	Var (Δp)	0.109	0.783	0.113	0.600	0.046	0.326	0.089	0.570
		(0.755)§		(0.686)§		(0.753)§			
	Var (Δ log p)	0.885	10.074	0.921	7.062	0.366	4.012	0.724	7.049
		(0.839)§		(0.771)§		(0.834)§			
Detroit:									
Premium	Var (Δp)	0.410	0.486	1.235	0.514	0.566	0.333	0.737	0.445
		(0.110)		(−0.501)‖		(−0.263)			
	Var (Δ log p)	3.179	6.376	10.427	5.787	4.523	3.445	6.043	5.202
		(0.429)#		(−0.343)		(−0.135)			
Regular	Var (Δp)	0.394	0.544	1.196	0.518	0.634	0.368	0.741	0.477
		(0.219)		(−0.502)‖		(−0.270)			
	Var (Δ log p)	3.899	9.168	12.496	7.644	6.329	5.491	7.575	7.441
		(0.539)§		(−0.299)		(−0.071)			
Los Angeles.									
Premium	Var (Δp)	0.200	1.123	0.037	0.977	0.117	0.944	0.118	1.015
		(0.702)§		(0.927)§		(0.784)§			
	Var (Δ log p)	1.546	11.008	0.256	10.348	0.831	9.522	0.878	10.293
		(0.756)§		(0.952)§		(0.843)§			

Regular	Var (Δp)	0.063	1.040	(0.897)§	0.069	1.022	(0.874)§	0.073	1.181	(0.888)§	0.068	1.081
	Var ($\Delta \log p$)	0.539	12.485	(0.924)§	0.641	12.852	(0.906)§	0.626	15.441	(0.925)§	0.602	13.593
New York												
Premium	Var (Δp)	0.044	0.227	(0.695)§	0.058	0.182	(0.584)§	0.034	0.138	(0.615)§	0.045	0.182
	Var ($\Delta \log p$)	0.276	2.013	(0.776)§	0.366	1.653	(0.703)§	0.215	1.266	(0.719)§	0.286	1.644
Regular	Var (Δp)	0.044	0.360	(0.785)§	0.050	0.213	(0.648)§	0.034	0.232	(0.749)§	0.043	0.268
	Var ($\Delta \log p$)	0.335	4.033	(0.849)§	0.386	2.473	(0.760)§	0.256	2.651	(0.826)§	0.326	3.052
Philadelphia												
Premium	Var (Δp)	0.031	0.025	(−0.120)	0.035	0.038	(0.046)	0.037	0.038	(0.027)	0.034	0.034
	Var ($\Delta \log p$)	0.231	0.276	(0.096)	0.259	0.403	(0.241)	0.273	0.411	(0.306)	0.254	0.363
Regular	Var (Δp)	0.032	0.030	(−0.030)	0.040	0.041	(0.017)	0.052	0.050	(−0.038)	0.041	0.040
	Var ($\Delta \log p$)	0.297	0.402	(0.161)	0.373	0.548	(0.200)	0.495	0.655	(0.232)	0.388	0.535

NOTE.—Prices deflated by BLS city price indices

* Reported values are actual values × 10^3

† See n. 19 herein

‡ (1) = January-April-July-October quarters, (2) = February-May-August-November quarters, and (3) = March-June-September-December quarters.

§ Variability of low prices is significantly greater than that of high prices at the 1 percent level

Variability of low prices is significantly greater than that of high prices at the 5 percent level

‖ Variability of high prices is significantly greater than that of low prices at the 1 percent level

TABLE 4

VARIANCES* OF QUARTERLY FIRST DIFFERENCES IN RETAIL GASOLINE PRICES (PITMAN'S TEST STATISTIC† IN PARENTHESES)

CITY AND GASOLINE GRADE	VAR (Δp)		VAR (Δ log p)	
	High Prices	Low Prices	High Prices	Low Prices
Atlanta ‡,§				
Premium	0.025	0.342	0.175	3.378
	(0.863)‖		(0.902)‖	
Regular	0.034	0.557	0.284	6.451
	(0.911)‖		(0.933)‖	
Baltimore ‡				
Premium	0.042	0.130	0.289	1.095
	(0.559)‖		(0.636)‖	
Regular	0.031	0.086	0.276	0.887
	(0.515)‖		(0.574)‖	
Boston #,§				
Premium	0.024	0.263	0.179	2.837
	(0.845)‖		(0.894)‖	
Regular	0.027	0.187	0.272	2.456
	(0.773)‖		(0.827)‖	
Buffalo**				
Premium	0.088	0.276	0.598	2.513
	(0.575)‖		(0.673)‖	
Regular	0.067	0.305	0.575	3.596
	(0.649)‖		(0.734)‖	
Cincinnati ‡				
Premium	0.036	0.344	0.254	3.545
	(0.811)‖		(0.868)‖	
Regular	0.038	0.333	0.332	4.367
	(0.799)‖		(0.863)‖	
Cleveland **,§				
Premium	0.023	0.078	0.163	0.639
	(0.608)‖		(0.651)‖	
Regular	0.022	0.078	0.195	0.792
	(0.563)‖		(0.608)‖	
Dallas **				
Premium	0.030	1.165	0.247	13.448
	(0.958)‖		(0.969)‖	
Regular	0.043	1.167	0.465	17.119
	(0.930)‖		(0.948)‖	
Honolulu †				
Premium	0.029	0.013	0.143	0.075
	(−0.550)††		(−0.510)††	
Regular	0.015	0.019	0.086	0.137
	(0.146)		(0.246)	
Houston †				
Premium	0.038	0.853	0.310	13.125
	(0.920)‖		(0.957)‖	
Regular	0.014	1.287	0.149	23.445
	(0.981)‖		(0.988)‖	
Kansas City †				
Premium	0.102	1.550	0.822	18.914
	(0.877)‖		(0.917)‖	
Regular	0.105	2.496	1.062	39.187
	(0.920)‖		(0.947)‖	
Milwaukee**				
Premium	0.485	0.835	3.956	9.505
	(0.282)		(0.434)‡‡	
Regular	0.655	0.801	7.104	11.523
	(0.124)		(0.290)	

TABLE 4—*Continued*

Minneapolis.#						
Premium	0.090	0.868	(0.839)‖	0.673	8.671	(0.883)‖
Regular	0.094	0.599	(0.753)‖	0.881	7.067	(0.807)‖
Pittsburgh #						
Premium	0.075	0.346	(0.644)‖	0.564	3.647	(0.732)‖
Regular	0.095	0.248	(0.496)‖	0.878	3.033	(0.603)‖
Saint Louis·†						
Premium	0.178	0.578	(0.556)‖	1.290	7.897	(0.739)‖
Regular	0.167	0.547	(0.569)‖	1.542	8.481	(0.724)‖
San Diego **						
Premium	0.109	0.477	(0.698)‖	0.781	4.262	(0.749)‖
Regular	0.118	0.591	(0.703)‖	1.052	6.504	(0.749)‖
San Francisco †,§						
Premium	0.051	0.536	(0.832)‖	0.352	4.850	(0.869)‖
Regular	0.073	0.558	(0.804)‖	0.632	6.515	(0.853)‖
Seattle **						
Premium	0.607	0.582	(−0.032)	4.842	5.574	(0.103)
Regular	0.625	0.593	(−0.038)	6.283	6.982	(0.076)
Washington **,§						
Premium	0.026	0.034	(0.131)	0.184	0.250	(0.161)
Regular	0.034	0.021	(−0.358)	0.288	0.204	(−0.258)

NOTE.—Prices are deflated by BLS city price indices.
* Reported values are actual variances × 10^3
† See n. 19 herein.
‡ March-June-September-December
§ BLS gasoline prices available monthly and BLS city price indices available quarterly
‖ Variability of low prices is significantly greater than that of high prices at the 1 percent level.
January-April-July-October
** February-May-August-November
†† Variability of high prices is significantly greater than that of low prices at the 1 percent level
‡‡ Variability of low prices is significantly greater than that of high prices at the 5 percent level

a random walk. Let p_t denote the price in period t. If the elements of the sequence $\{p_t\}$ follow a random walk, then

$$p_t = p_{t-1} + \varepsilon_t, \text{ where } E\varepsilon_t = 0,\ E\varepsilon_t^2 = \sigma_t^2, \text{ all } t$$

Consequently, var $(p_t - p_{t-1}) = \sigma^2$. For quarterly first differences, we have

$$p_t - p_{t-3} = \varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2}$$

Therefore, the variance of quarterly first differences, under the random-walk hypothesis, is $3\sigma^2$.

The evidence for low prices is much less consistent with the random walk in those cities where the variance of first differences of prices is relatively small. In other cities, for example, Los Angeles, variances of monthly first differences actually exceed those of the corresponding quarterly first differences. This indicates the presence of considerable negative serial correlation in these series.[22] Such a pricing pattern can be explained by the influence of stock holdings on prices. Prices must fall below normal if stock holdings are above normal; hence, the existence of short-term uncommitted stocks of gasoline in a particular market will drive prices down until such stocks disappear. I have already noted that "distress" or unanticipated stocks of gasoline are most likely to impact the low-price segment of the market These stocks will therefore serve to increase price variability in the low-price market relative to that in the high-price market The negative serial correlation arises from prices returning to "normal" levels, that is, the levels of prices prevailing if all marketers are relying for product supply on long-term supplier arrangements as stocks obtained on the spot market are exhausted

Another indication that high prices approach random-walk behavior more closely than do low prices is contained in the results reported in table 5 This table reports first-order autoregressions of the various BLS gasoline price series. The regressions are of the form

$$p_t = \alpha_0 + \alpha_1 p_{t-1} + \varepsilon_t$$

for those cities with monthly gasoline samples and

$$p_t = \alpha_0 + \alpha_1 p_{t-3} + \varepsilon_t$$

for cities sampled quarterly. We expect that, for a random walk, α_1 will be close to unity The values of α_1 contained in table 5 indicate that high prices almost always approximate a random walk more closely than do low prices The only major exceptions are the series for Detroit, but it will be recalled that for Detroit, high-price variability exceeds that of low prices Values of α_1 are generally closer to unity for those series with little variability of prices over time This suggests that there is a common (random) source of price variability in all series but that the low-price series incorporate additional sources of variability which do not affect the high prices We shall find that this additional source of variability has to do with local market conditions and is quite sensitive to the amount of information consumers choose to obtain.

Further evidence that the determinants of prices in the low-price segment of the gasoline market are substantially different from those of high prices is contained in table 6. This table contains correlations of the various BLS price series, on the one hand, and Platt's Oklahoma Group 3

---

[22] If the disturbances are, in fact, correlated, neglecting end-point effects, we have var $(P_t - P_{t-3}) = \text{var}\,(P_t - P_{t-1})(3 + 4r_1 + 2r_2)$, where r_i is the ith-order serial correlation coefficient Thus, var $(P_t - P_{t-3}) > 3$ var $(P_t - P_{t-1})$ implies that r_1 or r_2 (or both) is negative.

TABLE 5

FIRST-ORDER AUTOREGRESSIONS OF BLS GASOLINE PRICES

City	Values of α_1			
	Premium High	Premium Low	Regular High	Regular Low
Cities for Which Monthly Data Are Available $p_t = \alpha_0 + \alpha_1 p_{t-1} + \varepsilon_t$ (t Refers to Months)				
Atlanta	0.801	0.072	0.654	0.064
Boston	0.974	0.729	0.993	0.753
Chicago	0.957	0.652	0.964	0.659
Cleveland	0.998	0.960	0.997	0.960
Detroit	0.523	0.869	0.563	0.860
Los Angeles	0.806	−0.118	0.909	−0.064
New York	0.995	0.773	1.004	0.758
Philadelphia	1.000	0.973	0.991	0.963
San Francisco	0.962	0.553	0.925	0.552
Washington	0.981	0.982	0.989	0.987
Cities for Which Data Are Available Quarterly $p_t = \alpha_0 + \alpha_1 p_{t-3} + \varepsilon_t$ (t Refers to Months)				
Baltimore	0.927	0.793	0.950	0.759
Buffalo	0.963	0.884	0.980	0.832
Cincinnati	0.952	0.373	0.966	0.420
Dallas	0.943	0.032	0.950	−0.062
Honolulu	0.991	1.021	1.015	0.947
Houston	0.843	−0.091	0.941	0.099
Kansas City	0.864	0.160	0.868	−0.047
Milwaukee	0.552	0.270	0.488	0.130
Minneapolis	0.840	0.079	0.842	0.284
Pittsburgh	1.003	0.372	0.941	0.568
Saint Louis	0.926	0.501	0.933	0.595
San Diego	0.882	0.455	0.854	0.284
Seattle	0.754	0.623	0.722	0.585

regular grade gasoline prices, on the other. The Platt's price is the best available indicator of nationwide wholesale prices. A high correlation between the Platt's series and a BLS price series indicates that the stations whose prices comprise the BLS series are price takers responding to supply prices determined on the national level. The fact that the high prices are more highly correlated with the Platt's wholesale series than are low prices seems consistent with the view that an important determinant of the variability of low prices is the sale on a city-by-city basis of unanticipated stocks of excess gasoline at bargain prices by large suppliers.[23]

To this point I have compared price variability at low-price stations with that at high-price stations in the same city. The results in tables 2–4

---

[23] Note that this national market determination of the high prices would predict their random-walk behavior, for no one locality would be large enough to have an appreciable independent impact on such prices.

TABLE 6

CORRELATIONS BETWEEN THE BLS RETAIL GASOLINE PRICE SERIES AND PLATT'S OKLAHOMA GROUP 3 REGULAR GRADE WHOLESALE PRICES, 23 CITIES (JANUARY 1964–JANUARY 1971)

	Premium Grade Gasoline		Regular Grade Gasoline	
Metropolitan Area	High Prices	Low Prices	High Prices	Low Prices
Atlanta	0.630	0.117	0.554	0.000
Baltimore	0.588	0.606	0.658	0.554
Boston	0.626	0.045	0.620	0.339
Buffalo	0.667	0.688	0.624	0.674
Chicago	0.630	0.319	0.596	0.383
Cincinnati, Ohio, Kentucky, Indiana	0.572	0.387	0.526	0.314
Cleveland	0.643	0.564	0.648	0.564
Dallas	0.636	−0.107	0.611	0.135
Detroit	0.563	0.380	0.550	0.379
Honolulu	0.657	0.671	0.602	0.625
Houston	0.769	0.418	0.724	0.478
Kansas City, Missouri, Kansas	0.672	0.318	0.639	0.369
Los Angeles	0.657	0.168	0.695	0.531
Milwaukee	0.512	0.094	0.578	−0.018
Minneapolis	0.585	0.679	0.596	0.480
New York	0.630	0.415	0.528	0.407
Philadelphia, Pennsylvania, New Jersey	0.633	0.608	0.636	0.647
Pittsburgh	0.713	0.348	0.801	0.436
Saint Louis, Missouri, Illinois	0.662	0.659	0.664	0.691
San Diego	0.582	0.017	0.596	0.011
San Francisco	0.437	0.421	0.459	0.228
Seattle-Everett, Washington	0.500	0.479	0.563	0.448
Washington, D.C.	0.602	0.571	0.571	0.563

also indicate the existence of substantial differences in price variability across cities. For example, the amount of price variability exhibited by Los Angeles regular grade low prices is 10 times that of the corresponding Cleveland price series. My hypothesis relating price variability and information suggests that differences in price variability across cities will arise, in part, from differences by city in the incentives consumers have to obtain information about gasoline prices. To test this hypothesis, the following equation was estimated:[24]

$$\ln \operatorname{var} (\Delta \ln p)_i = \alpha_0 + \alpha_1 \ln b_i + \alpha_2 \ln y_i + \alpha_3 \ln s_i + \alpha_4 \ln m_i + \alpha_5 \ln u_i + \varepsilon_i, \qquad (2)$$

where the proxy variables for information and the market share variable are those described in Section 1 and u_i is the variance of annual unemployment rates by city included to account for the city-specific demand component of the uncertainty input to the gasoline market.[25]

---

[24] Prices used to construct the dependent variable have been deflated by BLS city indices of consumer prices. For those cities with data available on a quarterly basis, the dependent variable takes values reported in table 4. To allow comparability with these values, the averages of variances of quarterly first differences given in table 3 are used for monthly-data cities.

These regressions show once again the difference in behavior between prices at high-price stations and those at low-price stations. Since the relatively well informed customers make their way to low-price outlets, we would expect the information proxy variables to have the largest impact on the stations at the low-price end of the offer price distribution. This is clearly the case, for the coefficients of the information proxies are significantly different from zero and have the expected signs in the low-price regressions while those in the high-price regressions are insignificant. These results are consistent with my earlier hypothesis that high-price sellers are price takers from their suppliers and so charge prices which are effectively determined on the national level. Low-price sellers, on the other hand, are sensitive to the amount of search going on, with low prices being most variable in those cities with a relatively large proportion of well-informed customers. In this connection it is noteworthy that, while a larger fraction of the variance of the dependent variables across cities is explained in the low-price regressions, the standard errors of the estimates are almost equal. The table 7 regressions thus provide support for the position that high- and low-price variability both depend on a common set of exogenous influences but that low prices are also influenced by local market conditions. The strength of this influence depends on the amount of information consumers choose to obtain.

The coefficients of the other two explanatory variables, those designed to represent market structure and the variability of general economic conditions, show that these variables have an impact on the entire gasoline market, although in terms of the absolute amount of price variability resulting from a given increase in either variable, the effect is largest in the low-price market.[26] The fact that both of these variables are significant

---

[25] The unemployment rate data used are gathered by the BES from the state employment security agencies responsible for unemployment insurance programs. Such rates are therefore based on claims by unemployed workers for unemployment compensation. Because laws defining the coverage of such insurance programs differ widely across the states, it is difficult to make meaningful city-by-city comparisons of unemployment rates. However, by using the variance of the annual employment rates rather than the rates themselves, I was able to avoid many of the pitfalls of cross-city comparison. As long as the BES rates vary directly with true employment, my measure of variability will accurately reflect demand uncertainty, and since the data problems are in good part due to differences in coverage by state, we can expect measurement error to be consistently in one direction.

[26] The mean premium low-price variability (var $[\Delta \log p]$) is 0.00554, and that of premium high prices is 0.00104. At the means, a 1 percent increase in H leads to a decrease of 1.046×10^{-4} in premium low-price variability and a similar increase in the variance of the unemployment rate to an increase in 1.968×10^{-5} in premium low variability. The corresponding changes in the variability of premium high prices are, respectively, 2.214×10^{-5} and 9.142×10^{-6}. Thus, in absolute size, the effects of these variables on low-price variability are larger than their effects on the variability of high prices.

TABLE 7

PRICE VARIABILITY REGRESSIONS, 23 CITIES
DEPENDENT VARIABLE = LOG VAR (Δ LOG p)
PRICES DEFLATED BY BLS CITY PRICE INDICES

DEPENDENT VARIABLE PRICE SERIES	COEFFICIENTS (t-STATISTICS IN PARENTHESES)						
	Constant	Log Gallons per Car	Log Median Family Income	Log Median Years Schooling	Log Herfindahl Index	Log Variance of Annual BES Unemployment Rates, 1964–71	R^2 (SEE)
Premium high	−40.963	−0.456	0.660	9.377	−1.574	0.650	0.62
	(−1.93)	(−0.23)	(0.38)	(1.29)	(−2.39)	(4.23)	(0.76)
Premium low	−15.634	7.245	−4.717	20.611	−1.887	0.355	0.80
	(−0.80)	(3.99)	(−3.00)	(3.10)	(−3.13)	(2.53)	(0.70)
Regular high	−50.796	−0.739	0.761	12.458	−2.093	0.649	0.66
	(−2.37)	(−0.37)	(0.44)	(1.70)	(−3.15)	(4.18)	(0.77)
Regular low	−10.696	7.144	−5.710	22.342	−1.876	0.355	0.75
	(−0.47)	(3.38)	(−3.12)	(2.90)	(−2.68)	(2.17)	(0.81)

in the *high*-price regressions lends support to the view that the primary impact of consumer information on low prices is through encouraging low-price stations to participate actively in the spot market In addition, this added information serves to transmit shocks originating in the spot market throughout the low-price segment of the market The spot market, affecting only the low-price segment of the total gasoline market, is thus quite a different source of uncertainty than is variation in general economic conditions The variation in demand for gasoline in a particular locality is a diffuse source of uncertainty that relies only secondarily on consumers' information to transmit its broad direct impact and thus affects all stations in the market

The significant coefficient of the market structure variable in both the high- and the low-price regression does not allow one to infer that competition is impaired in those markets with sales concentrated in the hands of a few suppliers The coefficient may simply reflect the number of independent pieces of information which are being introduced to the market If suppliers act independently, the supply of uncertainty might be expected to increase as a greater number of independent decisions are taken The coefficient of market structure is also compatible with collusion Price variability may fall as concentration of sales in the hands of a few suppliers rises, simply because cheating on a collusive agreement is most likely when the collusive agreement is weak Such cheating, occurring sporadically, would be most likely to occur in low-concentration markets, while in high-concentration markets, those operating the collusive agreement may well find it easier to promulgate and enforce a stable price than a series of variable prices Unfortunately the information available does not allow one to distinguish between the competing hypotheses of competition and collusion

Table 8 repeats regressions of table 7, substituting dollar sales per car. The comparatively poor performance of these regressions is predicted by the theory of costly information I have pointed out that the gain from an additional unit of search depends on the dispersion of prices, not on their level Thus, the only effect of a higher price level on search is through a reduction in the number of units purchased—an effect which is already accounted for by the gallon consumption measures The sales measures therefore contain spurious variation across cities and so do not serve as well in the regressions as consumption measures It is interesting to note that the income coefficients in the low-price regressions in table 8 are smaller than the corresponding coefficients in table 7 This is because the income coefficient should properly be a partial derivative measuring search cost, but since the benefits from search are not held constant by sales measures, the income coefficient reflects some of the extra consumption of gasoline and so is forced to zero

The variances of first differences in prices, the other results serving

TABLE 8

PRICE VARIABILITY REGRESSIONS, 23 CITIES
DEPENDENT VARIABLE LOG VAR (Δ LOG p)
PRICES DEFLATED BY BLS CITY PRICE INDICES

DEPENDENT VARIABLE PRICE SERIES	COEFFICIENTS (t-STATISTICS IN PARENTHESES)						
	Constant	Log Sales per Car	Log Median Family Income	Log Median Years Schooling	Log Herfindahl Index	Log Variance of BES Annual Unemployment Rates, 1964–71	R^2 (SEE)
Premium high	−37.690	3.052	−0.225	13.153	−1.365	0.662	0.65
	(−1.84)	(1.23)	(−0.13)	(1.83)	(−2.63)	(4.49)	(0.73)
Premium low	−11.452	5.894	−4.619	19.111	−3.079	0.356	0.69
	(−0.47)	(1.99)	(−2.25)	(2.23)	(−4.97)	(2.02)	(0.88)
Regular high	−48.220	2.305	0.004	15.737	−1.857	0.659	0.68
	(−2.28)	(0.90)	(0.002)	(2.13)	(−3.47)	(4.32)	(0.76)
Regular low	−6.586	5.799	−5.609	20.849	−3.051	0.355	0.65
	(−0.24)	(1.77)	(−2.48)	(2.21)	(−4.47)	(1.83)	(0.96)

to characterize the price series, and the regression study all tell a consistent story. High-price stations operate in a market which is effectively separated from the influences governing prices at the low-price stations These two segments of the retail gasoline market differ in the amount of information that their customers choose to obtain, and therefore in the demand elasticities facing individual firms In addition, suppliers seem to behave differently toward the two types of stations The overriding point of this analysis is that consumer information stocks are an important determinant of price variability in the retail gasoline market This point should prove useful in studying price variability in other retail markets

References

Allvine, Fred C , and Patterson, J M *Competition Ltd The Marketing of Gasoline* Bloomington Indiana Univ Press, 1972

American Petroleum Institute *Petroleum Industry Factbook* Washington American Petroleum Inst , 1971

Cassady, Ralph, Jr *Price Making and Price Behavior in the Petroleum Industry* New Haven, Conn Yale Univ Press, 1954

Commerce Clearing House *State Tax Guide* New York Commerce Clearing House, 1967.

Duncan, A J *Quality Control and Industrial Statistics* 3d ed Homewood, Ill Irwin, 1965

Fisher, F M "Simultaneous Equations Estimation The State of the Art." *IDA Economic Papers* Arlington, Va Inst Defense Analysis, 1970

Learned, E P , and Ellsworth, C C *Gasoline Pricing in Ohio* Boston Graduate School Bus Admin , Harvard Univ , 1959

Livingston, S M , and Levit, T "Competition and Retail Gasoline Prices " *Rev Econ and Statis* 41 (May 1959) 119–32

Marvel, H P "Determinants of the Behavior of Retail Gasoline Prices An Application of the Economics of Information " Ph D dissertation, Univ Chicago, 1975

National Petroleum News "Mid-May Factbook Issue" (May 1971)

Rothschild, Michael "Models of Market Organization with Imperfect Information A Survey " *J P E* 81, no 6 (November/December 1973) 1283–1308.

Snedecor, G W , and Cochran, W G *Statistical Methods* 6th ed Ames Iowa State Univ Press, 1967

Stigler, G J "The Economics of Information " In *The Organization of Industry*, edited by G J Stigler Homewood, Ill Irwin, 1968

Telser, L G "Searching for the Lowest Price " *A E R Proc* 63 (May 1973) 40–49.

——— "Equilibrium Price Distributions " Unpublished Chicago Univ Chicago, 1974

U S Bureau of the Census Census of Business, 1967 *Retail Trade Gasoline Service Stations and Liquified Petroleum (LP) Gas Dealers* Publication BC 67-RS7. Washington Government Printing Office, 1971 (*a*)

———. Census of Business, 1967 *Retail Trade Merchandise Line Sales* BD67-MLS Washington. Government Printing Office, 1971 (*b*)

———. *State Tax Collections* GF71-1 Washington Government Printing Office, 1971. (*c*)

———. Census of Housing, 1970. *Detailed Housing Characteristics.* Final Report HC(1)-B23, Massachusetts. Washington: Government Printing Office, 1972. (*a*)

———. Census of Housing, 1970. *Housing Characteristics for States, Cities, and Counties.* Select numbers, vol. 1. Washington: Government Printing Office, 1972. (*b*)

———. Census of Population, 1970. *General Social and Economic Characteristics.* Final Report PC(1)-C, various numbers. Washington: Government Printing Office, 1972. (*c*)

U.S. Department of Labor. *Manpower Report of the President.* Washington: Government Printing Office, 1973.

A Theory of Government Enterprise

Cotton M. Lindsay

University of California, Los Angeles

A model of government enterprise is developed which contrasts the supply behavior of government and proprietary organizations. Refutable implications are derived from an explicit information structure in which the productivity of management must be monitored indirectly and differently for the two types of organizations Managers of both are assumed to be wealth maximizers It nevertheless is suggested that the output behavior of government enterprises will differ from that of proprietary enterprises in predictable ways—even where Congress intends the output of government bureaus to be identical to that of proprietary firms Numerous specific implications are developed and tested with reference to the operation of Veterans' Administration hospitals.

It is a widely held belief among laymen and economists that government organization of economic behavior produces different results from market organization of the same activity The title "bureaucrat," nominally indicating only membership in a government enterprise, has developed pejorative overtones reflecting these beliefs I suggest that the typical customer of a government bureau expects inconvenience and delay in receipt of service He is not surprised when the service is unavailable when demanded or defective when delivered, or when it is administered by personnel who are rude and indifferent

The author wishes to acknowledge the helpful comments of Tom Borcherding, Michael Darby, William Niskanen, Harold Demsetz, Kenneth Scott, and participants in the UCLA Law and Economics Workshop as well as the many perceptive and thoughtful suggestions of an anonymous referee A National Postdoctoral Fellowship of the Hoover Institution on War, Revolution and Peace facilitated completion of this work Robert Williams provided unusually careful and tenacious research assistance

[*Journal of Political Economy*, 1976, vol 84, no. 5]

© 1976 by The University of Chicago. All rights reserved

Widespread though these beliefs may be, economists have little but their own experience as customers to support these beliefs. Economic theory has yet to produce a satisfying counterpart to the profit maximizing firm with which to analyze public sector behavior. This paper seeks to fill this lacuna in the theory of economic organization by offering an explanation for these anomalies rooted in rational bureaucratic choice.

I

Where economic activity is organized by the market, interaction between demanders and suppliers occurs almost exclusively across the market. Buyers and sellers exchange offers and settle on terms. With democratic government organization, the customer communicates with the producer (in this case a government agency) through two channels. He continues his ordinary commerce with this supplier, but he also affects supply behavior through political channels. Influence through this second channel affects behavior in two distinct ways. First, individuals as a political group may demand from the supplying agency a different bundle of goods than they would purchase individually Secondly, because performance of government agencies is judged in terms of criteria other than profitability, bureaucrats are led to behave differently than personnel within proprietary organizations There are, in other words, both demand-based and supply-based reasons for government enterprises to behave differently than would proprietary firms confronting identical market demand functions. Although one rarely observes a government agency being influenced by only one side of this market in isolation, it is nevertheless useful to distinguish, at least conceptually, these two sets of influences.

Considerable attention has been devoted to the demand-based (i.e., political) influences on this behavior by economists and political scientists. For example, the familiar "free rider problem" discussed by Olson (1965) and others is alleged to suggest that output levels will be higher when political processes organize the provision of public goods than when private firms produce and sell those goods [1] Downs (1957), Black (1958), Riker (1962), Stigler (1972), and many others have developed theoretical models of the affect of parliamentary rules including rules affecting vote trading on governmental decisions and thus the allocation of resources in this sphere.

Less attention has been paid to the supply side of this market, and it is here that this paper will be focused. Although the belief is widespread that bureaucracies themselves influence the character of their output, the theory of why this is true is to date incomplete. The most popular approach to this problem is simply to postulate some alternative for profits as the agency manager's maximand Typically, the size of the bureau or its budget is assumed to be maximized.[2]

---

[1] See, however, Thompson (1968) and Lindsay (1972).

There are some obvious difficulties with this approach First it says too little about bureau behavior since, though managers are assumed to want larger bureaus, the theory does not directly address the manner in which these are obtained Budgets are typically treated as either exogenous or mechanically derived from congressional demand functions To say that bureaucrats have a taste for bureaus, therefore, tells us no more about their behavior as managers than does the assertion of any other taste. Even more critical to this approach, however, is the inconsistency of imputing to management the objective of maximizing their budgets when one of the most important functions of management seemingly is to husband resources of their organizations

Ultimately, all these difficulties stem from the failure of these presentations to model explicitly behavior on both sides of the market for managerial services Realistically, the size of the budget is of peripheral importance to both suppliers and demanders of these services It simply is not plausible that managers would exchange large amounts of money earnings for direct consumption benefits associated with budget size. Similarly, it cannot be the explicit intent of Congress to reward managers for simply spending money. Congressmen are interested presumably in pleasing their constituents, and this involves getting as much bureau output as they can for as little as they can Budget maximization "works" in explanations of bureaucratic behavior, therefore, only if informational constraints operating on both sides of the market cause increased spending by managers to be perceived as increased manager productivity by Congress [3] Explicit analysis of this market below will show this to be an implausible result Analysis of this market under realistic assumptions about how Congress monitors manager behavior does suggest, however, that there is a basis for a separate theory of government enterprise

II

It is assumed here that the demand prices for manager services in proprietary and government enterprises are related to the perceived productivity of their managers For proprietary firms the term "productivity" has a clear and unambiguous interpretation It measures a manager's marginal contribution to his firm's profits. Managerial productivity in a governmental enterprise is a concept which has enjoyed less currency in economic discourse. It is defined here to measure the contribution of these managers to social net income—that is, the difference between the social value of the bureau's output as determined by the political process and the cost of the resources used.

---

[2] See, e.g , McKean (1964), Parkinson (1962), Niskanen (1971), and Shapiro (1973)

[3] That the budget may be used as a proxy for output in judging the importance (i.e , productivity) of managers is a frequent rationalization of this rule. See, e.g , Shapiro (1973, p. 403)

If the social value of a bureau's output is the same as its market value, then the *actual* productivity of managerial behavior will not differ between competitive proprietary firms and government enterprises. Managers of these two types of enterprise, if rewarded on the basis of actual productivity, would make identical allocative decisions, and there would be no supply-based theory of bureaucracy distinct from the theory of the firm. Actual managerial productivity is not known in either case, however; owners of proprietary firms and members of Congress must discover the productivity of their respective managers. It is precisely because this information on manager productivity must be estimated and estimated differently in these two cases that behavior of the two types of organizations is itself predicted to differ.

Owners of proprietary firms gauge their executives' productivity by monitoring the financial statements of their firms. Clearly there are factors beyond the control of these managers which also affect firm profits. Study and comparison of the financial performance of their own and rival firms should nevertheless provide valuable and typically adequate information concerning the productivity of management. The discipline imposed by discerning customers, themselves searching for "good value," should make it unnecessary for these owners to monitor internal resource use or engage in quality control of their firms' products A manager who allows waste of inputs or unattractive product design will suffer in the first case a rise in cost and in the second a fall in price. Both will be reflected in his firm's profits and thus be noted as a decrease in his productivity by the ownership

Congress might also monitor profits, in which case again behavior would be indistinguishable between these two sorts of institutions At least this would be true as long as competition existed for both sets of enterprises. In many cases of government production, however, the social value of bureau output (or more specifically, the value placed on this output by the political process) is not the same as its market value. For example, the government provides ,housing, medical care, and education for the disadvantaged. The market value of such consumption is simply the amount these individuals will pay for these items. The very fact that the government chooses to provide them suggests that it regards this valuation, and therefore the amounts privately purchased, inadequate. The price charged this target population is therefore set politically at a level below cost. In such cases as these, profitability (total revenues minus costs) clearly does not correctly reflect real manager productivity Congress must attempt to monitor the productivity of government agency management differently.[4]

Lacking the convenient yardstick of profit, Congress has no choice but to attempt to estimate this productivity directly. That is, it must identify the influence of each manager on the net social product of his enterprise. To do this, obviously Congress must evaluate the output of such a government agency. The higher the difference between this evaluation and the bureau's budget, the more productive is presumed to be the manager.

In this evaluation process "quality control" in the product of the bureau is of great importance Managers organize production so as to maximize the perceived value of output, but it is Congress who determines what that value is. The value of bureau output is what Congress believes it to be worth. Congress may determine this worth, however, only by determining the levels of each of the various attributes associated with the product. There are certain attributes which Congress can monitor readily and others which it cannot Bureau managers are therefore influenced to divert resources from the production of attributes which will not be monitored to those which will In so doing they will increase the perceived value of the bureau's output As production of attributes not expected to be observed by Congress will not increase a manager's perceived productivity it will not pay for agencies to devote resources to their production

The conventional deterrent which operates to prevent proprietary firms from engaging in such quality depreciation, that is, a fall in demand price, is inoperative here Prices of government enterprises are typically politically determined (often set at zero), hence, insensitive to changes in consumer demand As long as Congress believes the value of this output is unchanged, then managers' earnings will remain unchanged

Consumers may, and often do, make known directly to Congress their dissatisfaction with the quality of bureaucratic output Information of this type may not be expected to replace effectively the information provided by falling market prices. Writing letters is costly and unlikely to provide tangible "relief" to the unhappy customer All but the most egregious shortcomings go unreported via this channel Second, the quality of the information provided is lower Individuals do not communicate to Congress the same sort of market validated information yielded by customers shopping around for bargains and expressing their satisfaction with purchases. In this respect, these comments resemble the tears and tantrums of small children which bear little correlation from moment to moment with the quality of their treatment by adults.

---

[4] The question of why government frequently chooses to organize this activity itself rather than finance its purchase in the market through a subsidy-voucher scheme cannot be addressed here. It is clear, however, that the bureaucratic problems identified in this paper would be eliminated by a shift to vouchers

We predict, therefore, that when proprietary and government enterprises confront identical demand functions, systematic differences in the behavior of these two types of enterprises will be observed The output of government enterprises will, in general, contain fewer of those attributes which are "invisible" to Congress, that is, whose presence and quantity are not monitored.

This may be made clearer with the aid of figure 1. The axes of this figure represent quantities of the two attributes of some product z. A production possibilities curve $t(z)$ identifies the combinations which may be embodied in each unit of the good with a given outlay. Curves $p^1(z)$ and $p^2(z)$ are isoprice curves linking combinations of attributes which will command a common price in the market As proprietary managers seek to obtain the most revenue for any given outlay, these isoprice curves are also congruent with managerial indifference curves Proprietary managers will thus be led to produce combination z^1 which maximizes the price obtainable along $t(z)$

Assuming that both z_1 and z_2 are monitored by Congress and that Congress's valuation of each attribute is the same as that revealed by consumers in the market, then government managers, too, will be influenced to produce combination z^1. Realistically, however, Congress does not meter the level of each of the thousands of attributes which describe these products As metering is costly and attributes vary in importance, it will not be worthwhile for Congress to monitor or even to sample the level of some attributes Occasionally, the cost of metering will be discontinuous, and it will pay to monitor some "threshold level" of an attribute but not be worthwhile to meter its presence above this amount Government manager indifference curves will in these cases resemble $\hat{p}^1(z)$ where attribute z_2 is perfectly invisible and $\hat{p}^2(z)$ where it is invisible beyond some threshold level. They will thus be influenced to produce z^3 in the former case and z^2 in the latter.

The theory presented thus suggests that government agencies will devote no resources to the production of invisible output. To the extent that Congress seeks to influence a bureau to produce a product which is no different from that provided by proprietary firms, such intentions will be disappointed in this respect. The output of such bureaus, containing nonoptimal combinations of product attributes, may be said to be of lower quality than output exhibiting attribute combinations in the "desired" proportions.

This is not to say that the level of certain attributes produced by government enterprises may not exceed the ideal level. Policemen whose

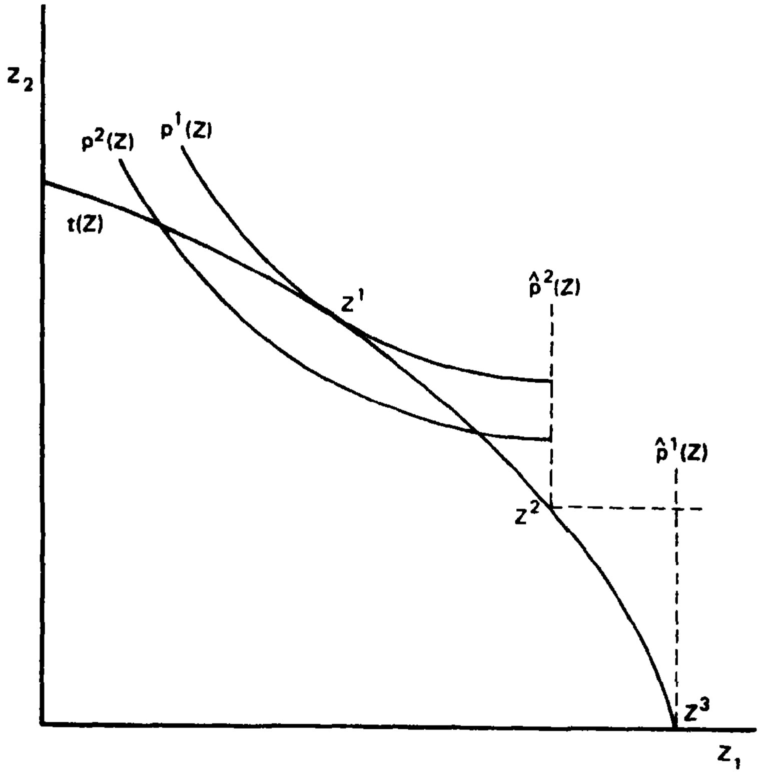


FIG 1

output is monitored by the number of summons which they issue may, in fact, patrol portions of road offering natural concealment more intensively than real traffic safety warrants The FAA may through regulations require that airline passengers purchase more safety from airframe failure (since airframe design is relatively cheap to monitor) than they would choose as part of their air travel package. Neither automobile nor air travel provides the utility which it might if the resources devoted by the Highway Patrol or the FAA to the uses mentioned were employed differently It is, therefore, convenient to use the phrase "lower quality" to denote the predicted phenomena of nonoptimal attribute combinations

By determining what sorts of output might be expected to be invisible, we may make even stronger predictions about the character of bureau output. Before turning to this task, let us consider, in the context of this "derived demand" view of managerial incentives developed here, an alternative view of the manner in which the productivity of managers is gauged

It is widely believed by noneconomists that salaries rise in bureaucratic hierarchies in relation to the "greater responsibilities" of higher office. An economic interpretation of this view holds that the larger is the bureau or the budget, the more productive is presumed to be (i.e , the more management services are supplied by) the manager In other words the levels of some or all *inputs* are used as a proxy for the value of manager services and these data alone enter in the manager's reward function Indeed, this may appear superficially to be a plausible interpretation of the ad hoc budget or bureau maximizing theories mentioned earlier [5] That is, if managers perceive their own earnings to be solely dependent upon the level of their budget or the size of their bureau, then such a variable becomes a logical maximand. The existing positive correlation between budgets and/or bureaus on the right-hand side and salaries on the left lends false credence to the idea that causation moves from right to left in this case

Such a scenario cannot describe the actual behavior of government enterprises No responsible government representative (or corporate director) would agree to compensate an executive on the basis of this type of reward function The informational content of such a proxy for managerial performance is nil, and it provides incentives which are positively perverse Managers confronting such a reward function have no reason (other than consumption) to produce *any* output and are rewarded for using up resources One imagines this process yielding a corp of bureaucrats supporting themselves in what used to be called "oriental splendor" by devoting government financed resources to the gratification of their own needs—*and being well paid for it* Unless some attributes of output are at least potentially meterable by Congress, *no* output will be produced

Aware of such a state of affairs, Congress will decline to fund output which it cannot meter. For the same reason we do not expect Congress to appropriate funds to finance the production of attributes which are invisible and therefore will not, in any event, be produced by government enterprises. This observation leads us to an important (and in the context of the literature on bureaucracy, unprecedented) implication of the model advanced here. The average cost of a government enterprise's output (being simply total cost divided by the level of some visible attribute) will in general be *less* than the average cost (calculated the same way) of a proprietary enterprise producing for the same market.[6]

---

[5] Niskanen has disavowed this interpretation in correspondence

[6] The behavior of government in this respect parallels that of the used automobile purchaser analyzed by Akerlof (1970) who expects to get a "lemon" and sets his demand price accordingly.

III

Two difficulties complicate efforts to test the theory empirically First, quality differences are predicted for enterprises facing identical demand functions. However, valid reasons often exist for believing that demand functions are not identical and that the political process *seeks* a lower quality product Where this is plausible, observation of low quality output is not particularly convincing evidence for the arguments presented here Indeed, were these two events always found in conjunction, the scientific value of this theory would be minimal As will be shown in section IV, implications may be developed from the theory which are not easily rationalized in terms of "demand side" phenomena

The second difficulty concerns the nature of the implications themselves Attributes which are predicted to be invisible to Congress might typically be expected to be invisible to an economist as well Predictions concerning the presence of attributes which are invisible to both are clearly difficult to verify directly Indeed, the theory seems, superficially at least, to be empirically empty almost by definition Fortunately, the theory may be made to yield additional implications which themselves support indirectly the hypothesis of inferior quality output For example, the suggestion at the end of section II that bureaus will produce output at lower apparent cost per unit itself *depends* on the presence of invisible attributes in proprietary firm output

Recent experience with the Veterans Administration hospital system serves both to illustrate applications of the theory and to provide empirical tests of its implications Table 1, for example, presents evidence concerning the cost hypothesis Reported cost figures for these two sources are not commensurable. The VA costs include the wages of salaried physicians, while physicians bill for their services separately in private hospitals Proprietary hospital costs, on the other hand, contain depreciation on buildings and equipment which is not included in VA costs Elsewhere (1975) I have estimated VA physician cost and proprietary depreciation cost per patient day for the years listed These have been subtracted from their respective totals in order to yield commensurable costs shown as "net costs" in table 1 As predicted, VA costs are lower than proprietary costs.

As is the case with most government enterprises, the chief source of information available to Congress concerning the character of its output is the VA's own statistical reports Reports from managers of VA hospitals and executives of the VA system are useful, of course, only if they can be expected to contain valid information This will be the case only if at reasonable cost the reports can be verified by audit. The cost of auditing varies greatly with the nature of the information to be verified, however. It seems plausible that factors affecting the feasibility of verifying the level

TABLE 1

VA AND PROPRIETARY HOSPITAL PER DIEM COSTS SELECTED YEARS

	VETERANS ADMINISTRATION			PROPRIETARY HOSPITALS		
YEAR	Reported	MD Cost	Net Cost	Reported	Depreciation	Net Cost
1969	$38.51	$12 29*	$26.22	$ 64 66	$ 6 39	$58 27
1970	46 44	12 91	33.53	76.80	8 97	67 83
1971	52 56	15.35*	37 21	87.25	11 84	75 41
1972	64 23	17.79	46.44	98 57	16.97	81.60
1973	69 32	20 23*	49.09	108.31	25 44	82 87

SOURCES —Committee on Veterans Affairs, U S House of Representatives, *Hospitals Guide Issues*, *1969–1973*, and *Operations of Veterans Administration Hospital and Medical Program*, various years, and author's computations

* Interpolated

of a particular attribute will include its value, its being witnessed by impartial parties, whether it is a continuing phenomenon or a brief event, whether it is a matter of routine or a singular episode Attributes which are of little value, are witnessed only by the supplier and the recipient, and are brief and episodal are predicted to be inauditable and therefore invisible to elected officials

One set of attributes of hospitalizations qualifying for our attention in this regard is patient service, whether rendered by personnel or capital equipment If the level of such services is reduced to the point where the incidence of death is detectably increased, then the attribute becomes, of course, quite visible Reductions in the levels of such characteristics above this threshold, merely impair the comfort of patients, however, and are extremely costly to monitor To the customer (patient) involved, neglect by staff, "crowding" of facilities, and absence of environmental amenities are quite noticeable and reason enough to avoid (and to advise friends and associates to avoid) patronizing such facilities in the future—if one is paying the bill Levels of such services, if reported, could not be effectively audited, hence such information is not sought by Congress The quantity of visits to the wards made by nurses and staff, the numbers of smiles and words of encouragement delivered, and the level of taste and thoughtfulness characterizing the patient environment in these hospitals thus remain invisible We may therefore predict that the level of such patient services per patient day is lower for VA hospitals than for proprietary hospitals [7]

---

[7] It will be noticed that the monitoring advantage of proprietary customers vis-à-vis Congress concerns the knowledge possessed by the former of specific "circumstances of time and place" pointed out by Hayek (1948, p 83) This advantage does not emerge, as Hayek claims, from the inability to commit information to statistics, since all information may with suitable manipulation be transformed into quantitative form. The principal difference lies simply in the cost of acquiring and validating the information at these two levels.

The cost of monitoring these characteristics is likewise high for us, but again indirect evidence supports this implication While we cannot ourselves (as Congress cannot) observe the level of personnel *services* provided, we can observe the number of staff employed Remembering that Congress would not be expected to fund invisible output attributes, we should expect them to support a budget for staff no larger than that required to provide the threshold (i e , life-maintaining) level of this attribute Congress would, in other words, have no interest in requiring the VA to devote funds to the employment of additional staff when it recognized that managers would have no pecuniary incentive to insure that this additional staff actually provided patient services Proprietary hospitals *do* have an interest in providing patient services above this "subsistence" level, however We therefore predict that staff/patient ratios are lower in VA hospitals than in proprietary hospitals Table 2 contains a comparison of mean staff/patient ratios for years when this data is available, VA staff/patient ratios are lower as predicted

Implications are also suggested concerning attributes exhibited by inputs to the hospital production process Personnel supplying services in this market may be said to possess in varying measure some attributes which are useful in the production of nonclinical services but are not useful in producing clinical services An example of such an attribute is "bedside manner " Among physicians with equal clinical abilities those with higher levels of these attributes will be able to command higher earnings in the market reflecting the market's evaluation of those attributes Only the output of clinical abilities of such personnel is visible to Congress Bedside manner, though valued by patients and perhaps by government representatives as well, cannot be metered by this latter group economically Managers of VA hospitals therefore have no incentive to expend the additional funds necessary to attract personnel with these attributes It is thus suggested that among personnel with equal qualifications (e g , physicians and nurses) those employed by the VA will earn less than those employed by proprietary hospitals or those who are self-employed

If most customers prefer to be served by white clerks, it also suggests that the Post Office and other customer service agencies will employ a disproportionately large number of minority members in customer contact positions Similarly we should expect the VA to employ a disproportionately large number of minority physicians If most hospital patients prefer to be treated by white physicians, then the attribute "white race" will command a positive premium in the market for physicians Managers of VA hospitals will find it not in their interest to pay this premium for this invisible attribute. They will therefore tend to hire disproportionately more minority physicians obtaining in the process the visible attribute—therapeutic services—at lower cost Table 3 presents

TABLE 2

STAFF/PATIENT RATIOS FOR VA AND PROPRIETARY HOSPITALS: VARIOUS YEARS

Year	VA General Hospitals	Proprietary Short-Term General Hospitals
1969	1.42	2.44
1970	1.46	2 56
1971	N/A	2.62
1972	N/A	2.67
1973	1.59	2 72

SOURCES.—Committee on Veterans Affairs, U S House of Representatives, *Operations of Veterans Administration Hospitals and Medical Program*, various years, and *Hospitals Guide Issues, 1969–1973*

TABLE 3

REPRESENTATION OF MINORITY PHYSICIANS ON VA HOSPITAL STAFFS 1970

	VETERANS ADMINISTRATION		TOTAL CIVILIAN U.S.	
	N	% Total	*N*	% Total
Total physicians	4,912	100 00	280,557	100 00
All minority	553*	11 26*	26,657	9 50
Spanish surname	251*	5 10*	10,451	3 72
Orientals	211*	4 30*	9,987	3 56
Blacks	89†	1.81†	6,044	2 15
American Indian	2†	0 04†	175	0 06

SOURCES.—U S *Census of Population*, 1970, and VA personnel records.
* Significantly different from U S proportion at 99 level
† *Not* significantly different from U S proportion at 90 level

evidence on minority representation among VA physicians in 1970. It was reported in 1970 that the proportion of VA physicians holding degrees from foreign medical schools was 25 percent while foreign-trained physicians represented 17 percent of all those licensed in the United States (U.S. Congress 1970). The former differs from the latter by approximately 15 standard deviations.

The invisibility of hospital care itself may explain the puzzling phenomenon of "excess utilization" of VA hospitals. A patient who is gravely ill requires a great deal of costly treatment, while a patient who has fully recovered or was never ill requires little. If "patient days" is the only attribute of output visible to Congress, then a given output may be produced with fewer resources if the patients filling the wards are of the latter type. If a VA patient who requires no care is replaced by one who requires some, then resources which might be devoted to the visible attribute, patient days (i.e., hotel services), must be diverted to the invisible attribute, care. Such substitutions reflect unfavorably on VA hospital manager productivity, hence we may expect them to make fewer than their proprietary counterparts. For example, we predict that, among similar individuals having experienced similar surgery, patient

TABLE 4

AVERAGE LENGTHS OF STAY IN VA AND VOLUNTARY HOSPITALS BY SURGICAL PROCEDURE FOR PATIENTS 50–64 YEARS EXCEPT WHERE NOTED
(REPORTED BY DAYS)

Procedure	VA	Voluntary
Pilonidal Cyst (35–49)	15.7	5.8
Diabetes mellitus	19.0	9.0
Acute coronary occlusion	31.5	21.7
Hemorrhoidectomy	15.3	7.1
Tonsils and adenoids (35–49)	6.4	2.4
Duodenum ulcer	15.2	6.7
Appendicitis	12.1	6.9
Inguinal hernia	17.1	7.2
Gastroenteritis and colitis	11.1	7.7
Gallstones	26.5	11.9
Pyelitis, cystitis, nephritis	11.6	6.0
Kidney stones	18.6	8.2
Prostate	22.1	9.7

SOURCE.—Interagency length-of-stay study group (unpublished).

TABLE 5

AVERAGE LENGTH OF STAY IN PROPRIETARY AND STATE AND LOCAL GOVERNMENT SHORT-TERM GENERAL HOSPITALS (IN DAYS)

	Proprietary	State and Local Government
1965	6.3	8.5
1966	6.4	8.6
1967	6.8	9.0
1968	7.0	8.9
1969	6.8	8.9
1970	6.8	8.7
1971	6.6	8.3
1972	6.6	8.0
1973	6.7	7.8
1974	6.7	7.7

SOURCE.—*American Hospital Association Guide to the Health Care Field*, 1975.

stays would be longer for VA patients than for patients of private hospitals. Table 4 reports some of the results of a comparison of VA and voluntary hospital patient stays in which age and cause of hospitalization are controlled.

It is widely alleged that VA patients are nontypical for reasons other than those outlined above. Table 5 therefore presents comparisons of lengths of stay for proprietary hospitals and all state and local government short-term hospitals. Although it is not possible to control here for procedure, the prediction toward longer lengths of stay in government hospitals is provided some limited support by this data.

Additional evidence of "excess utilization" was reported by the GAO in a recent audit of medical records of VA hospitals. It was discovered that 8.6 percent of the patient days furnished at these hospitals were performed on patients who could have been effectively treated in either nursing homes or as outpatients and who would not typically have been treated in non-VA hospitals. Another 5.9 percent of these days were furnished to patients awaiting surgery or undergoing tests typically performed on non-VA patients prior to admission (U.S. General Accounting Office 1973).

IV

We must interpret this evidence with care Long patient stays are seemingly as visible to Congress as they are to us and have been noted repeatedly by observers of the VA medical program. One would expect that this visibility of longer patient stays would prevent, or at least attenuate, the exploitation of such "false economies" It would seem that the perceived productivity of managers whose patients remained hospitalized for twice as long as patients in other hospitals would be adjusted downward for this phenomenon Yet longer patient stays persist.

Secondly, the private hospital sector is not free of its own anomalies affecting resource allocation Most hospital care purchased privately is insured, which suggests the possibility of moral hazard.[8] Consumers of insured services will demand more in both quantity and quality than they would purchase on a strict fee-for-service basis, and private hospitals may supply it.[9] The evidence presented here is not fully consistent with a moral hazard explanation either, however One may, on the one hand, interpret the lower quality of VA hospital service as an efficient response to customer wants which are not distorted by insurance-based demands for excess quality On the other hand, if moral hazard were the dominant influence here, we would expect a similar reaction in the quantity dimension, as well The observation of longer patient stays in VA hospitals is inconsistent with this implication

Finally, the implications concerning relative qualities of private and government output are not tested if Congress places a lower value on quality than do private demanders[10] It may be argued, in other words, that quality of government bureau output is low because Congress wants to provide low quality output. Lacking direct evidence of their intentions, we may only judge the importance of this objection by considering its plausibility in each case. This explanation may not be ruled out for the low quality observed in connection with VA hospitals. It may very well be the intention of Congress to provide a quality of care inferior to that typically supplied by proprietary hospitals in this country. The low staff/patient ratios, the large wards, the low-cost physicians may all be manifestations of the will of the political process to provide hospital care with a lower level of amenities than that which their patients would purchase for themselves The oft-repeated claim of the VA management that the VA provides a quality of care "second to none" need not be interpreted as seriously inconsistent with this view Evidence presented here concerning experience with VA hospitals should therefore be treated as merely suggestive

---

[8] See Pauly (1968)

[9] Feldstein (1971) and (1973) has argued that insurance will result in an increase in the quality of service supplied by nonprofit voluntary hospitals This is rationalized by giving hospital administrators a taste for quality which insurance allows them to gratify more fully. An alternative explanation of quality inflation associated with insurance is a simple variant on the moral hazard argument Reimbursable insurance influences hospitals to engage in quality competition As long as insurance covers costs, there is no reason for demanders or suppliers to economize on quality any more than they are influenced to economize on quantity in the more conventional moral hazard discussions (see Lindsay 1975, p 63).

[10] Inferior output is predicted for government bureaus even where Congress seeks to provide higher quality than the market.

It is not difficult to develop hypotheses and examples of behavior less easily rationalized in terms of political intent, however The intervention of government in a market is not typically coupled with a wish to depreciate the quality of the product supplied If public statements by government officials are to be believed, this intervention in many cases is principally motivated by a wish to improve quality For example, readers of the Beveridge Report, the blueprint of the English National Health Service, cannot interpret this document to portend the sort of gross degradation of the medical industry which has occurred there since 1948. Proprietary schools and universities are deplored by many educators because they feel that the quality of education in these schools would be substandard Yet Peltzman (1973) found that the net effect of government involvement in education was a dramatic decrease in the quality of the product provided Similar unsubstantiated statements abound concerning the low quality of medicine practiced at proprietary hospitals and the superiority of other organizational forms in this quality dimension

The preoccupation of monetary and fiscal authorities with the CPI rather than the real "cost of living" may be interpreted in this fashion. Price controls which increase the cost of living in invisible ways while holding down the highly visible CPI are seen to be attractive policy instruments The excessive sensitivity of the FDA to the possibility of carcinogens in our food and drugs is also understandable The visibility of an FDA screening failure such as thalidomide is clearly higher than the visibility of other costs imposed on consumers by more rigorous testing and finer screening, such as delays in development and more costly drugs. Herein may also lie the explanation for the universal practice of government-operated universities admitting only the "best qualified" students, that is, those whom it is least costly to teach rather than rationing space on the basis of income or price University managers are aware that the number of graduates and average graduate achievement level are regularly monitored attributes of university output. The oft-lamented output vector of university faculty as regards teaching and research is consistent with this model as is the familiar police "speed trap." Even the paradoxical affirmative action "quota" as a means of eliminating race- or sex-based discrimination becomes understandable.

The foregoing is not intended to suggest that government administrators are any less capable, less sensitive, or less dedicated public servants than their proprietary counterparts. Economics is a science in which differences in personality, ability, and taste play little part. Rather it is a reflection of the organizational scheme in which incentives, indeed survival characteristics, facing management leave little discretion. Increases in the quality of government enterprise output increase the "unit" cost of the product. The increase in the quality may not be apparent to elected officials The increase in cost almost certainly will be.

References

Akerlof, George. "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism." *Q.J.E.* 74 (August 1970): 484–94.

Black, Duncan. *The Theory of Committees and Elections.* Cambridge: Cambridge Univ. Press, 1958.

Downs, Anthony. *An Economic Theory of Democracy.* New York: Harper & Row, 1957.

Feldstein, Martin. "Hospital Cost Inflation. A Study of Nonprofit Price Dynamics." *A.E.R.* 61 (December 1971). 853–72.

———. "The Welfare Loss of Excess Health Insurance." *J.P.E.* 81, no. 2, pt. 1 (March/April 1973): 251–80.

Hayek, F. A. *Individualism and the Economic Order.* Chicago. Univ. Chicago Press, 1948.

Lindsay, Cotton M. "Impurities in the Theory of Public Expenditure." Discussion paper no. 21, Department of Economics, University of California, Los Angeles, 1972.

———. *Veterans Administration Hospitals: An Economic Analysis of Government Enterprise.* Washington: American Enterprise Institute, 1975.

McKean, Roland N "Divergencies between Individual and Total Costs within Government." *A.E.R.* 54 (May 1964): 243–49.

Niskanen, William A. *Bureaucracy and Representative Government.* Chicago: Aldine, 1971.

Olson, Mancur. *The Logic of Collective Action.* Cambridge: Harvard Univ. Press, 1965.

Parkinson, C. Northcote *Parkinson's Law and Other Studies in Administration.* Boston: Houghton Mifflin, 1962.

Pauly, Mark. "The Economics of Moral Hazard: Comment." *A.E.R.* 58 (June 1968): 531–36.

Peltzman, Sam. "The Effect of Subsidies-in-Kind on Private Expenditures. The Case of Higher Education." *J.P.E.* 81, no. 1 (January/February 1973): 1–27.

Riker, William H. *The Theory of Political Coalitions.* New Haven, Conn.: Yale Univ. Press, 1962.

Shapiro, David L. "Can Public Investment Have a Positive Return?" *J.P.E.* 81, no. 2, pt. 1 (March/April 1973): 401–13.

Stigler, George. "Economic Competition and Political Competition." *Public Choice* 13 (Fall 1972): 91–106.

Thompson, Earl. "The Perfectly Competitive Production of Public Goods." *Rev. Econ. and Statis.* 50 (February 1968): 1–12

U.S. Congress, House. *HUD-Space-Science Appropriations for 1971.* 92d Cong., 1st sess., 1970, p. 440.

U.S. General Accounting Office *Better Use of Outpatient Services and Nursing Care Bed Facilities Could Improve Health Care Delivery to Veterans* Report to Congress by the comptroller general of the United States, April 11, 1973.

Exhaustible Resources and Industrial Structure: A Nash-Cournot Approach to the World Oil Market

---

Stephen W. Salant

Board of Governors of the Federal Reserve System

The theory of exhaustible resources is modified to take account of the industrial organization of the world oil market. The cartel is viewed as a unified enterprise which dominates other extractors because of its larger reserves. Equilibrium price and sales paths are derived giving neither the dominant extractor nor the competitive fringe any incentive to change its intertemporal behavior. Under standard but simplified cost assumptions, it is shown that a disproportionate share of the increased profits resulting from the formation of the cartel goes to nonmembers and that the cartel's restriction on sales eventually leaves it the sole supplier of oil.

The current structure of the world's oil industry bears little resemblance to the extremes assumed in the theoretical literature on exhaustible resources. There is neither a single cartel (or firm) which owns all the world's oil and thus has unchecked power to set prices over time, nor is there an abundance of measureless, "Mom-and-Pop" oil extractors dotting the globe. Instead, the industry contains one cartel with more power than any other extractor, but other extractors do exist and have enough importance, collectively and perhaps individually, to restrain the full exercise of monopoly power.

In this paper, the conventional theory of exhaustible resources is modified to take account of this structure. Since many extractive industries besides oil have market structures "intermediate between monopoly and

The views expressed herein are solely those of the author and do not necessarily represent the views of the Federal Reserve System. I am grateful for the insights of Jeff Shafer and Steve Salop in formulating this problem. I also wish to thank—without implicating—Richard Berner, Michael Dooley, Dale Henderson, Peter Isard, Harry Johnson, Val Koromzay, Sam Peltzman, Michael Salant, and Martin Weitzman.

[*Journal of Political Economy* 1976, vol. 84, no. 5]
© 1976 by The University of Chicago. All rights reserved

perfect competition"—as Hotelling (1931, p. 171) remarked long ago[1]—the modified theory may also be applied to other exhaustible resources.

The world oil industry is regarded as a collection of independent firms from which consumers purchase oil at the same price Each firm may be viewed as a participant in a Cournot noncooperative game—maximizing its own discounted profits while taking as given the sales path of the remaining firms. Assumptions about the extraction costs and oil reserves of each firm determine whether the Cournot equilibrium which results resembles more closely monopoly, competition, oligopoly, or price leadership.

As a simplification, every small extraction unit or "plant" is assumed to have the same oil stock and cost function In the absence of a cartel, each plant is owned by a different firm A cartel is said to form when more than one plant comes under control of the same firm The cartel is then in the distinctive position of owning larger reserves than any other firm Moreover, if the marginal cost function of each plant is upward sloping, the cartel can extract at the same rate as another firm for a smaller cost

These two distinguishing characteristics give the cartel market power over every other firm If it is assumed that the rest of the world's oil stock is divided equally among a sufficiently large number of other extractors, a dominant firm model results In such a case, each small firm acts like a competitor which takes as given the price path set by the cartel and chooses a sales path to maximize the sum of discounted profits. The cartel takes the sales path of the "competitive fringe" as given and chooses a price path—supported by cartel sales—to maximize its discounted profits Any situation where each sector takes as given the optimal choice of the other and where neither can, under that assumption, increase its profits by altering its own strategy is called a Nash-Cournot equilibrium [2]

This paper analyzes equilibrium properties of the dominant extractor model [3] In the next two sections, it is assumed that costs at each plant are a linear function of the rate of extraction (constant marginal costs). As a result, the acquisition of several plants by the cartel does not enable it to extract more cheaply than another firm Nonetheless, the cartel still has one important advantage over the other extractors. it owns more oil. To focus on this advantage, which has no counterpart in the theory of cartels selling an ordinary producible, marginal extraction costs are assumed to be zero [4] Section I considers the optimization problems of both the cartel and the competitive fringe under this cost assumption, and then shows how the equilibrium price and sales paths can be constructed Qualitative properties of the resulting equilibrium are discussed in section II and comparisons are made to the more familiar extremes of monopoly and perfect competition In section III, the dominant extractor model is reconsidered under the assumption that the marginal cost of extracting at each plant is upward sloping

---

[1] Although Hotelling recognized the importance of oligopoly in extractive industries, he did not analyze the case in detail Since his pioneering article, theoretical research on exhaustible resources has focused exclusively on the rare extremes of monopoly and perfect competition

[2] The equilibrium is named after Cournot and Nash because the first originated and the second generalized the assumption that each player ignores the effects of his actions on the strategies of the others. In contrast, Stackelberg assumed that one particular player can manipulate the reactions of the other The unexplained asymmetry of this alternative approach is itself a drawback; moreover, it makes an extension of the analysis from duopoly to oligopoly impossible.

[3] The versatility of the Cournot approach is further illustrated in Appendix B, where asymmetric oligopoly is shown to result when different assumptions are made about stocks owned by the various extractors The dominant extractor model of the text is then shown to be a limiting case of asymmetric oligopoly.

Three results emerge under either cost assumption (1) the formation of the cartel raises the discounted sum of profits of the competitive fringe by an even greater percentage than those of the cartel, (2) after the initial jump in prices, the cartel continues to raise the real price of oil monotonically over time, and (3) no matter how small its supplies compared to those of the competitive industry as a whole, the cartel restricts its sales so as to take over the market after the competition has exhausted its reserves

I. The Oil Cartel as a Dominant Extractor: The Case of Zero Marginal Costs

To begin, consider the problem facing the cartel Suppose the cartel knows the stationary consumer demand curve, the sales path of the competitive sector, and the stock of its own resources The sales of the competitive sector may include sales of competitive speculators who begin with nothing but are free to enter, purchase in any period, and sell a smaller or equal amount later The cartel's problem is to pick a path of prices, supported by its sales, to maximize its discounted revenues Alternatively, it can be thought of as picking its sales path over time

Since the cartel takes the competitive sales path as given, it can deduct the given sales on each period from the consumer demand curve to obtain a sequence of excess demand curves Each curve will indicate the amount of oil demanded of the cartel at a given price in excess of what the competitive fringe sells at that instant Faced with these excess demand curves, the cartel must choose a sales path which maximizes the sum of discounted revenues, but which does not exceed its initial inventory

---

[4] To extend the results of the next two sections (and the appendixes) to the case of positive, constant marginal costs, reinterpret the words "price," "marginal revenue," and "choke level" as "net of the constant marginal cost "

To accomplish this, the cartel (like a monopolistic extractor) must compare the marginal revenue associated with selling another unit today to the discounted marginal revenue of selling it in the future The marginal revenue associated with a sale by the cartel of Q^m units is simply the price it would receive for selling another unit less the loss it would incur, because of the slight reduction in price, on the inframarginal units (Q^m) it sells·

$$M = P(Q^m + \bar{Q}^c) - Q^m \cdot a(P),$$

where

Q^m is the sales of the cartel,
$\bar{Q}^c$ is the (given) sales of the competitive sector,
P is the price at which consumers would purchase $Q^m + \bar{Q}^c$ units,
M is the marginal revenue associated with the excess demand curve, and
$a(P)$ is the absolute value of the slope of the consumer demand curve (with respect to sales), expressed as a function of price

Notice that this marginal revenue is derived from the *excess* demand curve faced by the cartel

If the strategy of the cartel is optimal, it must generate marginal revenues of the same discounted value in all periods in which the cartel sells, once the cartel stops selling, the discounted marginal revenue must never exceed the marginal revenue on any period of positive sales. Violation of either condition would provide opportunity for additional profits Hence, along the optimal path, the marginal revenue grows at the rate of interest until sales cease; then it grows at a smaller (or equal) rate Cumulative sales along the optimal path must, in addition, equal the cartel's initial inventory (I^m), leaving some oil in the ground is never optimal

Given the competitive sales path, the cartel picks its sales path to maximize discounted profits Alternatively, it can be viewed as selecting the optimal price path, supported by its sales The price path it selects may have any shape, since the excess demand curves it faces in successive instants depend on the arbitrarily given competitive sales path The price path may have regions where the price rises by more than the rate of interest or even falls absolutely. However, we must ask if all other agents in the model would react to the chosen price path in the way assumed by the cartel. If not, the price path does not result in noncooperative equilibrium To understand how the other agents will react, the competitor's problem must be considered.

The competitive sector takes the price path as given and knows its total stock. It chooses a sales strategy to maximize discounted profits without exceeding its initial inventory (I^c). Since the competitive sector contains speculators who may *buy* for re-sale—as well as extractors—its "sales" may be negative (i.e., purchases) in some periods.

The competitor's optimization problem, like the cartel's, is entirely conventional. The competitors will always exhaust their stock. As long as the price rises by the rate of interest, different ways of selling a given stock have the same discounted value. If the price rises by a greater amount between two periods, the competitive speculators would enter and attempt to make *unlimited* profits by buying in the first period and selling in the next. If the price begins to rise by less than the rate of interest, the competitive sector will find selling its inventory before this region is reached to be optimal.

These two maximum problems provide enough information to characterize the equilibrium when it exists. Price rises at the rate of interest as long as competitors hold stocks. Afterward, it can rise at a smaller or equal rate. Price can never rise by more than the rate of interest in equilibrium. The marginal revenue derived from the excess demand curve must rise at the rate of interest, as long as the cartel makes positive sales, when they are completed, the marginal revenue can rise at a smaller or equal rate. Each sector ultimately sells its entire stock.

These conditions imply that, as long as the cartel begins with positive stocks, it will conserve enough to continue selling after the competitors drop out.[5] For, suppose the cartel completed its sales before the competitors. Then the price path would rise at the rate of interest while the two sectors coexisted and would continue to rise at that rate after the cartel stopped selling. Now, compare some early moment when the cartel is selling to some later moment when its sales are zero. When its sales are positive, its marginal revenue will be less than the price, when its sales are zero, its marginal revenue will be equal to the price (since it has no inframarginal units on which to take losses). Since price is growing at the rate of interest, the marginal revenue would have to grow by more, but this would give the cartel an incentive to alter its strategy. Hence, in equilibrium, the competitors cannot continue selling after the cartel drops out.

If both sectors begin with positive stocks, they must both operate simultaneously in the market for a period of time. When the price reaches some level, the competitive fringe completes its sales and abandons the market to the cartel. Denote this "termination" price as P^*. Prior to the time P^* is reached, the two sectors coexist in the market; afterward, the cartel operates alone. This suggests a solution technique. We begin at P^* and the marginal revenue at that price associated with the consumer demand curve, $MR(P^*)$. Before P^* is reached, the price and marginal revenues grow at the interest rate; hence, u moments before termination of the first phase at P^*, the price $P(u, P^*)$ and marginal revenue $M(u, P^*)$ are:

---

[5] In Section III, this result is shown to hold in a dominant extractor model with increasing marginal costs at each plant. In Appendix B, the result is shown for an oligopolistic market where the cartel owns more oil than any other single firm.

$$P(u, P^*) = P^* e^{-ru}, \qquad \text{and} \qquad M(u, P^*) = MR(P^*)e^{-ru}.$$

The price and marginal revenue, u moments before termination at P^*, each depend on two variables: the sales of each sector at that time:

$$P(u, P^*) = P[Q^m(u, P^*) + Q^c(u, P^*)].$$

$$M(u, P^*) = P[Q^m(u, P^*) + Q^c(u, P^*)] - Q^m(u, P^*) \cdot a[P(u, P^*)].$$

Substituting, we obtain:

$$P^* e^{-ru} = P[Q^m(u, P^*) + Q^c(u, P^*)], \tag{1}$$

$$MR(P^*)e^{-ru} = P^* e^{-ru} - Q^m(u, P^*) \cdot a(P^* e^{-ru}) \tag{2}$$

These two "marginal" equations can be solved simultaneously to obtain the sales of each sector u moments before termination at P^*: $Q^c(u, P^*)$ and $Q^m(u, P^*)$. Only these sales paths will permit price and marginal revenue to grow at the rate of interest until they reach, respectively, P^* and $MR(P^*)$

When the price reaches P^*, the competitors drop out and the cartel takes over the entire market. In this second phase, its marginal revenue must continue to grow at the rate of interest until demand is eliminated. Define the choke price (F) as the vertical intercept of the stationary demand curve. The marginal revenue begins at $MR(P^*)$ and grows at the rate of interest until it reaches F. We denote the cumulative sales of the cartel in the second phase as $\Delta(P^*)$

The duration of the first phase (S) and the price (P^*) at which the competitors drop out are determined by two "exhaustion" equations:

$$\int_0^S Q^c(u, P^*)\, du = I^c. \tag{3}$$

$$\int_0^S Q^m(u, P^*)\, du + \Delta(P^*) = I^m. \tag{4}$$

Equation (3) states that P^* and S must be chosen so that the competitors sell their entire inventory during the first phase; equation (4) states that P^* and S must be chosen so that the cartel sells its entire inventory during the two phases.

A sufficient condition for the existence of a noncooperative equilibrium (with nonrandomized strategies) is that the consumer demand curve have a point of unit elasticity ($\underline{P}$) and that elasticity along the curve increase

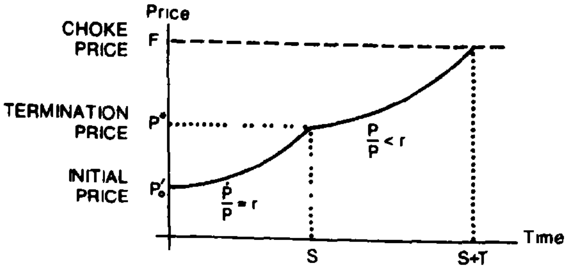


FIG 1 —The equilibrium price path in the dominant extractor model

strictly with price The proof is relegated to Appendix A All linear[6] or concave demand curves and many convex curves satisfy this sufficient condition

Whenever an equilibrium exists, the price path will have the appearance as shown in figure 1 As is indicated in the figure, the price begins at some level, P'_0, grows for S periods at the rate of interest, and then grows more gradually for T additional periods until the choke price is reached

During the entire first phase, price and marginal revenue remain in the same proportion For this to occur, the elasticity of excess demand (η^{ex}) must, in equilibrium, be the same at each point chosen by the cartel If the consumer demand curve satisfies the sufficiency condition mentioned above, the point elasticity of a *given* excess demand curve increases as the price rises So the competitive sales must fall as price rises in order to *shift* the excess demand curve to the right and restore the elasticity of excess demand to its previous value The elasticity of aggregate demand (η^{ag}),[7] that of excess demand multiplied by the cartel's market share, rises during the first phase with the cartel's share [8] Since the sales of the competitive fringe diminish as the first phase progresses, so does the flow of its discounted profits In contrast, sales and therefore the discounted profit flow of the dominant firm may increase during the first phase.[9]

---

[6] E g , in the case of the linear demand curve $P = F - a(Q^m + Q^c)$, equations (1) and (2) imply $Q^m(u, P^*) = (1/a)(F - P^*)e^{-ru}$ and $Q^c(u, P^*) = (F/a)(1 - e^{-ru})$ Since, in this case, competitive sales are independent of the termination price, (3) uniquely determines the duration of the first phase, (4) may then be used to determine the termination price of that phase

[7] Since the cartel always receives a positive marginal revenue, it always operates in the elastic region of the *excess* demand curve it faces However, the *aggregate* demand curve may be inelastic early in the first phase A failure to distinguish these two elasticities has led some to infer, merely because estimated aggregate demand for oil is inelastic, that OPEC's pricing policy is suboptimal

[8] In principle, therefore, data about market shares of a dominant extractor could be used to update any previous estimate of the elasticity of aggregate demand obtained during the first phase

[9] Since $\dot{Q}^m \gtreqless 0$ as $(P/a)(\partial a/\partial p) \lesseqgtr 1$, the sales of the cartel will increase as the first phase progresses unless the aggregate demand curve is very convex.

II. A Comparison of Different Market Structures

Consider the transition from a competitive world of many small extractors to a world where some of these extractors form a collusive cartel which dominates the remaining small firms. In figure 2, the equilibrium price path (A) prior to the formation of the cartel is portrayed. Price rises from its initial level (P_0) at the rate of interest until the world inventory ($I^m + I^c$) is exhausted, then the price remains at the choke level (F) Since selling a barrel of oil in any period has the same discounted value, each firm (including prospective members of the cartel) receives a total discounted profit of P_0 multiplied by its initial stock.

If the firms owning I^m barrels decide to collude while the owners of the remaining I^c barrels continue to act competitively, the equilibrium price path changes to that described in the previous section The new path (B) rises at the rate of interest during the first phase and at a slower rate during the second Moreover, since cumulative sales along the new path must remain equal to the unchanged world inventory, the two paths must cross This can happen only if the new price path begins above P_0 (at P_0'), rises at the rate of interest during its first phase and then crosses the competitive path in the second phase as indicated in figure 2

The formation of the cartel raises the sum of discounted profits of each extractor[10] However, the profits of every competitive firm jump by an *even larger percentage* than those of the cartel. This result arises here for exactly the same reason as in the theory of cartels selling an ordinary producible good The competitive firms get the full benefit of the higher prices without themselves having to contort their sales in order to generate those prices

To verify diagrammatically that the competitors get a "free ride," note that discounted profits of the competitive sector change from P_0I^c to $P_0'I^c$ when the cartel forms Profits of the firms which join the cartel, however, rise by a smaller percentage since the cartel must sell some of its stock during the second phase at discounted values below P_0'.

This situation, of course, gives individual members of the cartel an incentive to "chisel" or defect *provided* the others would maintain the cartel discipline For, if only a single member broke ranks, it could earn P_0' per barrel by selling all its oil before the second phase If, however, its chiseling led to the defection of the other members, the price would drop to the competitive level (P_0) and all extractors would be worse off Threats of quitting may be used to affect the division of proceeds within the cartel. However, members presumably recognize that maintenance of the cartel is in their collective interest. As Stigler has put it, "a large

---

[10] In standard welfare analyses of extraction, it is shown that only the intertemporal allocation associated with the competitive price path is Pareto-optimal. Since the presence of a dominant firm benefits all extractors, consumers must be the only losers.

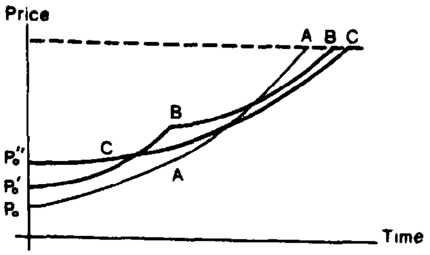


FIG. 2.—Comparison of price paths under different market structures

group of free riders will find that the street car won't run."[11] In the case of a firm (or cartel) selling an ordinary producible, another reason for the possible loss of a dominant position is that entrepreneurs attracted by the profits of the competitive fringe learn how to produce the good and eventually enter the industry. In the case of an exhaustible resource, however, mere knowledge of the extraction technology does not permit entry. The entrepreneur cannot enter without acquiring an additional amount of the resource, an impossibility in a model without new discoveries.

If all the firms were to join the cartel, the monopoly path (C) would result. Since cumulative sales along this path must equal those along the other two, the monopoly path must cut each of the others. For this to occur, it must begin above the other two (at P_0''), cut each of them in the first[12] phase, and terminate last. Pure competition may be regarded as one extreme of the dominant extractor model where the cartel owns virtually nothing. Similarly, pure monopoly may be viewed as the other extreme where the cartel owns virtually everything. In the former case, the entire path consists of a first phase, while in the latter case, the entire path consists of a second phase. The dominant extractor model is intermediate between the extremes of pure monopoly and pure competition in the sense that, given the *same* world stock, the price lies below the high monopoly price, but above the low competitive one, and, with a dominant firm, the fixed world stock is exhausted more rapidly than under monopoly, but less quickly than under competition.[13]

---

[11] See Stigler (1966, p. 233) for a discussion of cartels and free riders producing an inexhaustible flow good.

[12] The price path of the monopolist cannot cross the path of the dominant firm in its *second* phase. For if the two paths touched there, the marginal revenues would then be equal. And, since both would be growing at the rate of interest, they would remain equal. But this implies that the two price paths would have to stick together rather than cross.

[13] If the dominant firm initially owned a smaller amount of the fixed world oil stock, it can be shown that the price path would change in a predictable way provided the path were uniquely determined by the initial stocks. The more the competitive sector owned, the lower would be the initial price and the more quickly the world's oil would be depleted. The first phase of each lower path would cut the second phase of each upper path.

III. Extensions to the Case of Increasing Marginal Costs

If the marginal cost of extracting at each plant is upward sloping, not only does the acquisition of several plants by the cartel provide it with a larger stock of oil than any other firm, but also it enables the cartel to extract at the same rate for a lower cost Specifically, when the oil stock of the cartel is X times that of any other firm, the cartel will be able to produce at X times the rate of another firm for the same marginal cost That is when

$$I^m = \left(\frac{I^c}{n}\right) X,$$

$$\phi'(Q^m) = C'\left(\frac{Q^m}{X}\right),$$

where $C'(q^c)$ is the marginal cost of extraction at each of the n competitive plants, and $\phi'(Q^m)$ is the marginal cost of extraction for the cartel

Prior to the formation of the cartel, each firm acts competitively, taking the price path as given and maximizing discounted profits. The optimal extraction policy is one where the difference between price and marginal cost is increased at the rate of interest until supplies are exhausted. Since each firm is identical, all extract at the same rate at any instant and earn the same overall discounted profit If the firms which will comprise the cartel are considered as a group, they optimally extract at X times the rate of an individual firm at any instant and exhaust their stocks (which are X times as large) at the same moment [14] Prior to the formation of the cartel, the group earns X times the discounted profits of a typical firm

Once the cartel forms, the equilibrium[15] shifts If the cartel is very large relative to any other extractor, it sets the price path for the competitive fringe. In equilibrium, the excess of price over marginal cost at each competitive firm must grow at the rate of interest until supplies at each firm are exhausted; afterward, it grows at a smaller rate Similarly, the difference between marginal revenue (to the excess demand curve) and marginal costs of the cartel must grow at the rate of interest until its supplies are exhausted As will be shown, the equilibrium will have the same basic properties as in the case of constant marginal costs

First, any equilibrium must again involve a coexistence phase followed by a phase where the cartel takes over To prove this, note that once either sector operates alone, the other will never wish to sell later. If the competitive firms operate alone, price less their marginal cost must rise at the

---

The three paths in figure 2 can be considered as special cases from this entire field of price paths. Such a transfer of the world's oil stock from the cartel to the competitive sector would actually occur if some members of the cartel defected

[14] A concise analysis of the case where competitive firms have different initial stocks and extraction costs has been conducted by Goldsmith (1974).

[15] In contrast to the case of constant marginal costs, the general existence of equilibrium has not yet been formally established for the case of increasing marginal costs.

rate of interest. But then, the marginal profits of the cartel $[P - \phi'(0)]$ will rise at a slower rate, making it suboptimal for the cartel to have waited to sell in the future. Similarly, if the cartel operates alone, marginal revenue less marginal cost must rise at the rate of interest. But then price less the competitor's marginal cost $[P - C'(0)]$ will grow at a smaller rate,[16] making it suboptimal for a competitive extractor to have waited to sell in the future. Therefore, since it is optimal for both sectors to exhaust their initial stocks, any equilibrium must involve a coexistence phase followed perhaps by the operation of one sector.

Moreover, the competitive extractors cannot be the ones to operate last. If the competitive firms continued to sell after the cartel dropped out, then at the termination of the first phase, the marginal cost of each firm would exceed that of the cartel. To satisfy the first order conditions, marginal revenue less the marginal cost of the cartel would also have to exceed price less the marginal cost of each competitive firm at *all* previous moments. However, for the marginal cost of each competitive firm to exceed that of the cartel at every previous instant, each competitive extractor would have to operate at more than $1/X$th the rate of the cartel until the cartel exhausted its supplies—an impossibility, since each firm has only $1/X$th the stock of the cartel. Thus, in any equilibrium it is the cartel which must outlast the competitive extractors.

Second, as in the case of constant marginal costs, the real price of oil increases monotonically. For this result, it is sufficient that the cost and demand functions be stationary. For—assuming stationarity—suppose that the real price ever failed to rise between one period and the next. Then, for price less marginal cost to rise at the rate of interest, the output at each competitive firm would have to fall (assuming stationarity of costs). Similarly, for the marginal profits of the cartel to rise, its output would also have to fall. But, if the output of each sector fell, the price would (assuming stationarity of demand) rise—contradicting the assumption.

Third, a comparison of the competitive and dominant extractor equilibria reveals that—even in this more general case—the formation of the cartel raises the discounted profits of each competitive firm by an even greater percentage than those of the cartel.[17] Each firm initially earns $1/X$th the profits of the group which forms the cartel. When the cartel forms, the price path changes. The competitive firms devise a new sales path to maximize the value of their output along the new price path. If the new price path were fixed, the maximizing strategy for the cartel would be to extract in each period X times as much as each competitive extractor. Given the new price path, this strategy would restore the profits of the cartel to X times the profits of each firm. Of course, this strategy is infeasible since the price path would not then remain fixed, and so the cartel cannot choose this unconstrained optimum. The profit it gets from its optimal choice constrained to generate the new equilibrium price path will, therefore, be less than X times that of the typical competitive firm. Thus, when the cartel forms, a disproportionate share of the increase in discounted profits goes to the competitive fringe.

---

[16] For this proposition to be valid, it is sufficient to assume that elasticity of demand increases with price along the demand curve.

[17] The corresponding analysis in a Cournot model of a dominant firm selling an ordinary producible good may be found in Shubik (1959, pp 134–35).

Appendix A

Existence of Equilibrium

Equilibrium in the dominant firm model occurs when, given the price path of the cartel, the competitors cannot make greater discounted profits by selecting a different, feasible sales path, and, given the sales path of the competitors, the cartel cannot make greater profits by altering its price path Any solution with the following five properties is, therefore, an equilibrium for the case of zero marginal costs.[18]

1. As long as both sectors sell, price and marginal revenue grow at the rate of interest.
2. After the competitors drop out, marginal revenue continues to grow at the rate of interest until the cartel stops selling; it then remains at the choke level (F)
3. After the competitors drop out, price never grows faster than the rate of interest
4. Over the entire horizon, the cumulative sales of each sector exactly match its initial inventory.
5. At no intermediate point does the partial sum of sales of either sector exceed its initial stock (the requirement for a feasible sales path).

A condition sufficient for the existence of equilibrium is that the consumer demand curve have a point of unit elasticity ($\underline{P}$) and that elasticity of demand increase strictly with price

To prove that this condition is sufficient for the existence of equilibrium given arbitrary initial stocks, we show (*A*) If this condition holds, *any* solution to equations (1)–(4) has the five properties above (*B*) If this condition holds, there will *always* exist a solution to equations (1)–(4).

A. By construction, any solution to equations (1)–(4) will possess properties 1, 2, and 4. Furthermore, the condition mentioned above insures that any solution to (1)–(4) will have property 3 It must be shown, however, that any solution to (1)–(4) satisfies property 5 if the sufficiency condition holds. This follows since (1)–(4) implies a *nonnegative*[19] sales path for each sector and each sector ultimately exhausts its initial supplies.

---

[18] These require trivial modification for the case of constant marginal costs

[19] That $Q^m(u, P^*) \geq 0$ follows from equation (2) The proof that $Q^c(u, P^*) \geq 0$ relies on the sufficiency condition. Since

$$\eta^{ex} = \frac{(Q^m + Q^c)}{Q^m}\,\eta^{ed}(P),\ \frac{\partial \eta^{ex}}{\partial P} = 0 \qquad \text{for } \frac{\partial \eta^{ed}}{\partial P} > 0 \text{ only if } \frac{Q^m + Q^c}{Q^m}$$

decreases as P increases This requires that Q^c/Q^m fall monotonically as the first phase progresses. Since $Q^c(0, P^*) = 0$, competitive sales must be strictly positive at each moment of the first phase and must decrease as that phase progresses.

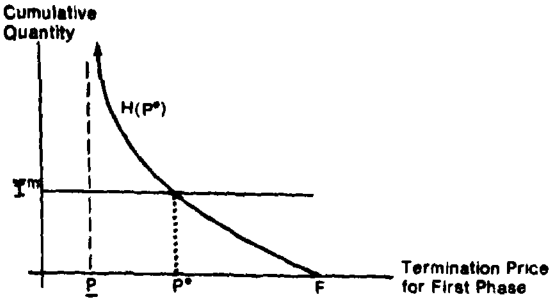


FIG 3 —Determination of P^*

B. Next, we show that there will always exist a solution to equations (1)–(4) provided the sufficiency condition holds Since (1) and (2) can always be solved for the sales function of each sector, what is required is a proof that a (P^*, S) combination always exists satisfying the two exhaustion equations. Since competitive sales are positive and decrease as the first phase progresses, there exists for *any* termination price (P^*), a unique duration (S) for the first phase satisfying equation (3) [20] Hence, equation (3) implicitly determines a positive function $S(P^*)$

Substituting $S(P^*)$ into (4) converts it to a single equation in one unknown (P^*)

$$H(P^*) = \int_0^{S(P^*)} Q^m(u, P^*)\, du + \Delta(P^*) = \bar{I}^m \tag{5}$$

A solution to this equation always exists since $H(P^*)$ is a continuous function equal to zero at the choke price and approaching infinity as P^* approaches $\underline{P}$ (the point of zero marginal revenue) from above.[21] Somewhere in between lies a termination price solving the exhaustion equations Figure 3 illustrates the determination of P^*

Appendix B

The Dominant Extractor Model as a Limiting Case of Asymmetric Cournot Oligopoly

Assume the oil industry to be composed of firms in two sectors extracting oil at zero cost and selling it to consumers at the same price. One sector consists of the cartel, which acts in the coordinated manner of a single firm Each of the n oligopolists in the second sector is assumed to own an equal fraction of the sector's oil. None, however, owns as much as the cartel $[\bar{I}^m > (\bar{I}^c/n)]$

---

[20] If $Q^c(u, P^*) \leq 0$ for some termination price, there might exist no duration which would exhaust the competitive stock The same problem might arise if competitive sales, although positive, decreased toward zero as we moved back further from termination of the first phase In either case, the function $S(P^*)$ might not exist

[21] $H(F) = 0$ since $\Delta(F) = Q^m(u, F) = 0$

$$\lim_{P^* \downarrow \underline{P}} H(P^*) = \infty \quad \text{since } Q^m(u, P^*) \geq 0, \qquad \text{while } \lim_{P^* \downarrow \underline{P}} \Delta(P^*) = \infty$$

—the cumulative sales are infinitite along a path which begins where marginal revenue is *zero* and continues forever with the marginal revenue "growing" from that point at the rate of interest

Every firm in the oil industry (including the cartel) is assumed to take the aggregate sales path of the other n firms as given and, taking account of the effects of its own sales on price, to devise its optimal, intertemporal sales strategy. The price path is determined by the aggregate sales paths of the $n + 1$ extractors. The characteristics of the optimal strategy for each extractor (including the cartel) are that: (1) the marginal revenue rises at the rate of interest until the firm exhausts its supplies, and then rises at an equal or smaller rate; and (2) each firm exhausts its initial inventory.

Since the marginal revenue of either type of firm is simply the price it would receive for selling an additional unit of oil less the revenue it would lose (because of the resulting slight reduction in price) on the inframarginal units it was selling, the firm selling the larger number of inframarginal units has the lower marginal revenue.

At the end of the first phase, the marginal revenue of the type of firm which drops out will exceed that of the firm selling to the entire market. From the marginal conditions, it is clear that this same relationship in marginal revenues must hold at any moment of the coexistence phase. This implies that the sales of the firm destined to drop out are smaller than the sales of the firm which will take over at each moment of the first phase. This, in turn, implies that, in equilibrium, the cartel cannot be the first to exhaust its supplies

For, suppose the cartel dropped out first Then, regardless of the length of the first phase, the marginal conditions would imply that cumulative sales of the cartel would be smaller than cumulative sales of each smaller firm. But then the exhaustion equations would never be satisfied, since any first phase long enough to exhaust the cartel's supplies would more than exhaust the stock of each smaller extractor. Since assuming the cartel is the first to drop out leads to an inconsistency between the marginal equations and the exhaustion equations, no equilibrium will have this characteristic In any equilibrium the cartel will continue to sell after the smaller oligopolists have exhausted their supplies The following equations must then hold.

$$P(u, P^*) = P[nq^c(u, P^*) + Q^m(u, P^*)], \tag{6}$$

$$P^*e^{-ru} = P(u, P^*) - q^c(u, P^*) \; a[P(u, P^*)], \tag{7}$$

$$MR(P^*) \; e^{-ru} = P(u, P^*) - Q^m(u, P^*) \; a[P(u, P^*)], \tag{8}$$

$$\int_0^S q^c(u, P^*) \, du = \frac{\bar{I}^c}{n}, \tag{9}$$

$$\int_0^S Q^m(u, P^*) \, du + \Delta(P^*) = \bar{I}^m. \tag{10}$$

Equation (7) states that the marginal revenue of a representative smaller firm u moments before it exhausts its supplies at P^* is equal to its discounted marginal revenue at termination. Equation (8) states that, u moments prior to termination of the first phase, the marginal revenue of the cartel will equal the discounted value of its marginal revenue when the n smaller firms depart.

The new system (6)–(10) is very similar to the model (1)–(4) studied in sections I and II. Except for notation, the only difference is between (1) and (7). The smaller firms no longer drive the current price down as far as the discounted value of the termination price because they consider the losses on their inframarginal units. Hence, price rises at less than the rate of interest during the first phase of the oligopoly model.

As the number of firms (n) dividing the noncartel oil stock grows, however, the sales of an individual firm at each moment of the first phase become negligible. If the slope of the demand curve is bounded, the last term of equation (7) vanishes in the limit and our system (6)–(10) converges[22] to (1)–(4) Thus, the price taking behavior of the competitive fringe emerges as the limiting case of the Cournot oligopoly model

References

Goldsmith, Oliver Scott "Market Allocation of Exhaustive Resources" *J.P.E.* 82, no 5 (September/October 1974) 1035–40

Hotelling, H "The Economics of Exhaustible Resources" *J P E* 39, no. 2 (April 1931) 137–75

Shubik, Martin *Strategy and Market Structure* New York Wiley, 1959

Stigler, George. *The Theory of Price* New York Macmillan, 1966

[22] For the case of the linear demand curve of n 6, eqq (6)–(10) imply that

$$nq^c(u, P^*) = \frac{nF(1 - e^{-ru})}{a(n + 2)}$$

and

$$Q^m(u, P^*) = \frac{F + n(F - P^*)e^{-ru} - (2P^* - F)e^{-ru}}{a(n + 2)}$$

As $n \to \infty$, these sales functions converge to those obtained in the dominant firm case of n 6

Communications

The Effect of Broadcasting on Political Campaign Spending: An Empirical Investigation

Burton A. Abrams and Russell F Settle

University of Delaware

With enactment of the Federal Election Campaign Act of 1971 and the amendments of 1974, Congress moved to resolve, at least partially, the perennial "problem" of escalating campaign spending by federal candidates The extent of the problem is suggested by the facts on presidential campaign spending presented in table 1 While there has been long-term growth in reported national-level spending on presidential elections (the annual growth rate over the 1868–1972 period was about 5 percent), the rapid escalation in campaign expenditures is a relatively recent phenomenon. Between 1948 and 1972, reported spending grew at an annual compound rate of approximately 10 percent, or, in real terms, at about 7–8 percent Even allowing for increases in the size of the voting-age population (see table 1), presidential campaign spending has grown substantially over the past several elections

What explains this dramatic growth in campaign spending? A widely accepted explanation is that increased reliance on radio and television advertising has produced an "arms race" in political campaign spending. For example, as early as 1952—the very first year that television played a significant role in political campaigns—television was blamed for stimulating "excessive" campaign expenditures[1] This belief that in-

---

We are indebted to George Stigler and to referees of this *Journal* for a number of useful suggestions Of course, the usual caveat applies. The order in which the authors' names appear was determined by a random selection process.

[1] See, e.g , "House to Test Campaign Spending· Parties Call TV Bane of Limitations," *New York Times* (December 1, 1952)

[*Journal of Political Economy*, 1976, vol 84, no. 5]
© 1976 by The University of Chicago All rights reserved.

TABLE 1

PRESIDENTIAL GENERAL ELECTION CAMPAIGN EXPENDITURES (IN DOLLARS), 1868–1972

YEAR	EXPENDITURES BY PARTY		EXPENDITURES PER ADULT
	Republican	Democratic	
1868	150,000	80,000	0.01
1872	250,000	50,000	0.01
1876	950,000	900,000	0.08
1880	1,100,000	360,000	0.06
1884	1,300,000	1,400,000	0.09
1888	1,350,000	860,000	0.07
1892	1,700,000	2,350,000	0.11
1896	3,350,000	680,000	0.10
1900	3,000,000	430,000	0.08
1904	2,100,000	700,000	0.06
1908	1,660,000	630,000	0.04
1912	1,070,000	1,130,000	0.04
1916	2,440,000	2,280,000	0.08
1920	5,420,000	1,470,000	0.11
1924	4,020,000	1,110,000	0.10
1928	6,260,000	5,340,000	0.16
1932	2,900,000	2,250,000	0.07
1936	8,890,000	5,190,000	0.17
1940	3,450,000	2,780,000	0.07
1944	2,830,000	2,170,000	0.06
1948	2,130,000	2,740,000	0.05
1952	6,610,000	5,030,000	0.12
1956	7,780,000	5,110,000	0.12
1960	10,130,000	9,800,000	0.18
1964	16,030,000	8,760,000	0.22
1968	25,400,000	11,600,000	0.31
1972	35,180,000	18,480,000	0.38

SOURCE.—1868–1968, Alexander (1972, p 6), 1972, Committee on Commerce (1973, p 322)

NOTE.—Expenditures are reported national-level spending on presidential general elections in current prices. Per capita expenditures were computed by dividing voting-age population into total presidential campaign expenditures by both parties.

creased reliance on the broadcast media heightened the level of total campaign spending received influential support from Heard's authoritative study, *The Costs of Democracy* (1960, p. 403). "Radio and television broadcasting by national-level campaign groups not only further displaced other media of publicity but had the *additional effect of increasing total costs*" (emphasis added).

Further support for this view came from the Committee for Economic Development (1968). In a study of election-financing reform, the committee (p. 16) found that "there are several legitimate reasons for the continuing escalation of campaign costs. Primary among these are increased reliance on radio and television, growing use of advertising and public relations services, outlays for travel, and rising prices in all these fields." A Ralph Nader Congress Project (Green et al. 1972, p. 9) was even more selective in placing the blame for higher election costs: "More than any other factor, television was the cause." Finally, Alexander (1972, p. 12) alleged that "television, jet travel, and polling are some of the *obvious reasons* that campaign costs have increased dramatically in recent years" (emphasis added).

In light of such claims, it is not surprising that the Federal Election Campaign Act (Public Law 92–225) attempted to restrain campaign spending by limiting the amount that federal candidates could spend on campaign advertising, especially broadcast advertising.[2] The 1971 act limited expenditures on such advertising forms as radio, television, cable television, newspapers, magazines, outdoor advertising, and automated telephone systems to 10 cents per voting-age person in the election area and, in addition, restricted spending on radio and television advertising to not more than 60 percent of the total allowable advertising expenditures. Moreover, the broadcast and the print media were prohibited from charging candidates more than the lowest unit-rate charges for commercial advertising Such provisions as these suggest that Congress also blames the mass media—especially radio and television—for fostering "undesirable" increases in campaign spending

In 1974 Congress went even further in its efforts to curb federal campaign spending and the various abuses of the democratic process that such spending has ostensibly fostered In amendments (Public Law 93–443) to the 1971 Campaign Act, Congress removed the ceiling on broadcasting expenditures but imposed a number of other ceilings on federal election campaign expenditures and on individuals' contributions to federal candidates However, in January 1976 the Supreme Court held that many of the expenditure and contributions limitations established by the 1974 amendments were unconstitutional [3]

While the argument that radio and television advertising stimulates excessive campaign spending enjoys a wide and long-standing acceptance, it is nevertheless questionable on both theoretic and empirical grounds. In principle, the interjection of broadcasting into political campaigns has an indeterminate effect on total campaign spending The introduction

[2] Some elected officials no doubt support campaign spending limitations on the grounds that such restrictions are likely to make it more difficult in future campaigns for relatively unknown but well-financed opponents to get their "message" across to the voters. Additionally, spending limitations serve to burden the entrance of new parties into the political arena

[3] To illustrate, candidates for Senate and House seats are—as a result of the Court's ruling—no longer subject to any spending limits Presidential candidates who do not accept federal funds are also free to spend as much as they can raise. For presidential candidates who do accept such funds, the maximum allowable expenditure remains at $10 million for the primaries and $20 million for the general election. The Court also upheld the provision limiting individual contributions to not more than $1,000 to any one candidate ($2,000 if both a primary and a general election are involved), although contributions to all candidates may total to $25,000 However, individuals may, under the Court's ruling, spend any amount in "independent" efforts to elect or defeat a candidate. For a complete discussion of the Court's decision, see "High Court Upholds Public Funds for Presidential Election Races, Removes Most Limits on Spending," *New York Times* (January 31, 1976).

of radio and television advertising represented technological advances in campaigning and, thus, produced a reallocation of campaign budgets to accommodate the purchase of broadcast advertisements. However, the effect of such technical change on the *level* of campaign spending is less clear, depending on potentially complex game-strategic responses and counterresponses by each candidate. To illustrate this theoretic indeterminacy, independent efforts by candidates to increase their share of the vote could escalate spending in response to the introduction of broadcasting, while collusive agreements among candidates to avoid "unprofitable" undertakings could reduce total campaign costs when technical change lowers unit (per vote) cost.[4] This "small-numbers" indeterminacy in "political markets" is similar, of course, to the well-known ambiguities associated with oligopolistic markets.[5] Since there is no satisfactory a priori basis for predicting even the direction of broadcasting's effect on political campaign spending, the issue must be resolved empirically

Previous studies (noted above) have done little more than demonstrate that over the past several decades campaign expenditures have grown pari passu with candidates' reliance on radio and television advertising. These investigations—relying on simple correlations—clearly neglect the possibility that, in addition to broadcasting, several other factors may have influenced the level of campaign spending. For example, growth in the size of the voting-age population, increases in the level of national income, or expansion in the opportunities for providing "legislative profits" as a quid pro quo for contributions could all heighten political campaign spending. In the present paper, we investigate the alleged relationship between radio and television usage and the level of campaign spending within a multivariate context that allows us to control for such nonbroadcasting influences.

Estimates of Broadcasting's Impact on Campaign Expenditures

In testing the widely accepted hypothesis that radio and television are responsible for much of the growth in political campaign expenditures over the last several decades, we estimated a number of models by ordinary least squares with time-series data on presidential general-election campaigns. In this section we discuss the various determinants of campaign spending and explain the operational forms of the variables used in quantifying the influence of broadcasting on total campaign expenditures. We then turn to a discussion of the empirical findings. While we make no pretense to having a fully specified empirical model, we do believe that our estimates have allowed us to learn something useful about the quantitative significance of the broadcast media for political campaign spending. We conclude with some comments on the implications of our findings for public policy toward campaign spending, especially spending on the broadcast media.

---

[4] Of course, these outcomes are merely illustrative and do not rule out other possibilities E.g , the introduction of broadcasting may lower unit (per vote) costs sufficiently to offset the expansionary influence of independently acting candidates attempting to maintain or increase their share of the votes. Additionally, the presence of broadcasting, which facilitates the transmission of information to voters, increases the likelihood that a candidate's support for special-interest groups will be revealed to voters, e.g., by an opposing candidate, thereby raising the cost to the candidate of supporting such groups. This effect could serve to attenuate the expansionary bias of maintenance-of-market-share expenditures. Interestingly, expenditures that serve simply to maintain market, or vote, share are socially wasteful in an economic efficiency sense; however, if such expenditures also stimulate voter interest and awareness and increase voter turnout, it is no longer quite so obvious in a democracy that these outlays are "wasteful."

[5] For further elaboration see, for instance, Scherer (1970).

A. *The Variables*

Presidential campaign spending.—For the dependent variable, presidential campaign expenditures by party, we relied on reported national-level committee spending on presidential general-election campaigns in constant 1957 dollars[6] (Alexander 1972, Committee on Commerce 1973, p. 322). While there are some difficulties with this variable (e.g., reported spending allegedly understates actual spending), we accepted it as a reasonably good, and certainly the best available, proxy for actual campaign expenditures. For each election year there are two observations for this spending variable, one for each party's national committees.

Radio and television.—As noted above, conventional wisdom informs us that, as radio and then television became more and more commonplace in the homes of potential voters, candidates began placing greater emphasis on broadcast campaigning, thereby touching off an "arms race" in total campaign spending. To capture this growth in the importance of the broadcast media, we measured radio usage as the proportion of all households with radios and television usage as the proportion of all households with televisions (Television Digest, Inc. 1973). The reasoning behind this formulation of the broadcast variables was that, as the proportion of radio (television) households increased from zero to nearly unity, radio's (television's) potential for substantially influencing total campaign spending increased *pari passu*.

We also experimented with binary variables as an alternative measure of the influence of broadcast usage. These variables assumed a value of zero for those elections occurring prior to the year that radio (television) became an important campaign instrument and a value of unity for all other elections.[7] This specification attempts to capture shifts in behavioral relationships between the preradio (pretelevision) and the postradio

---

[6] The wholesale price index was used to translate all current monetary magnitudes into constant 1957 dollars.

[7] Radio and television became important campaign tools in 1928 and 1952, respectively; see Heard (1960) for further discussion.

(posttelevision) periods. Since the parameter estimates were not sensitive to the particular measure of broadcast usage employed, we report (in the next section) only the estimates based on continuous measures of broadcast usage.

Income.—In addition to broadcast usage, a potentially important determinant of campaign contributions and, thus, spending is contributors' income. Assuming campaign contributions to be a "normal" good, we would expect secular increases in contributions and spending as real income grows, *ceteris paribus*. As a measure of real national income, we used GNP in 1957 prices (U.S. Bureau of the Census, various issues and 1966).

Voting-age population.—Another potentially important nonbroadcasting determinant of campaign spending is the size of the voting-age population, although on a priori grounds, the effect of a *ceteris paribus* increase in the size of the voting-age population on campaign contributions and expenditures is indeterminate. On the one hand, many contributions are probably made out of a sense of "good citizenship," implying that an increase in the number of potential voters will tend to increase total contributions as the new prospective voters strive to be good citizens. On the other hand, as Tullock (1967, p. 110) notes, population growth diminishes the potential influence of any one voter (contributor), thereby discouraging voting (contributing) (U.S. Bureau of the Census, various issues and 1960).

Federal spending.—The volume of campaign contributions and, therefore, spending also depends on the magnitude of the "legislative profits" that contributors stand to gain (e.g., if their candidate wins). Consequently, changes over time in the availability of opportunities for providing legislative profits to contributors would yield changes in the level of contributions and spending. As government involvement in economic affairs has become more extensive, additional legislative profits have presumably become available, especially to organized special-interest groups capable of lobbying for their positions and backing their efforts with substantial campaign contributions. As a control for the availability of legislative profits, we employed federal government expenditures in constant 1957 dollars[8] (U.S. Bureau of the Census 1960; Council of Economic Advisors 1974). We assume that this variable provides a rough indicator of the extent of governmental involvement in, and influence over, the economy, and thus of the opportunities for rewarding campaign contributions. A positive sign is anticipated for the coefficient of this variable.

Party affiliation.—Throughout the analysis we controlled for party affiliation with a dichotomous variable indicating whether a candidate was a Republican or a Democrat.[9] We accepted this variable as a useful proxy for some of the nonbroadcasting determinants of political campaign spending Republican candidates, for example, are commonly viewed as "pro-business" while Democrats are frequently seen as "pro-labor," and such distinctions—whether actual or imagined—could well affect the level of contributions and spending

---

[8] To account for the possibility that legislative profits have grown more rapidly than federal spending, we also experimented with the square of federal spending as a proxy for legislative profits. This transformation did not, however, enhance the explanatory power of our estimating equations.

Voter participation rate —We also included the voter participation rate as a control for several nonbroadcasting determinants of campaign spending (U.S Bureau of the Census, various issues and 1960) To illustrate the pertinence of this variable, in any one election a charismatic candidate or a closely contested race can heighten voter interest, which presumably manifests itself as increased voter turnout and campaign contributions (and, thus, campaign spending) In addition, secular growth in the availability of legislative profits or perhaps in the public's educational-attainment level may heighten voter interest, thereby stimulating both larger voter turnouts and greater campaign contributing and spending[10] Inclusion of the participation rate in the estimating equations allows us to control at least partially for such nonbroadcasting influences on the level of political campaign spending[11] A positive sign on the coefficient of this variable is anticipated

Admittedly, there may be reciprocal causation between voter interest (proxied by the participation rate) and campaign spending. The greater the voters' interest in an election, the more they are likely to contribute, and thus the higher will be spending However, the greater the expenditures on campaigning, the greater will be voter interest—assuming that campaign spending is "productive " Consequently, there is likely to be some simultaneity bias in our OLS estimates An additional possible complication is that the introduction and usage of broadcasting may be largely responsible for the higher levels of voter interest, which, in turn, generate more contributions and spending[12]

---

[9] Recall that each election provides two observations—one for each party.

[10] In fact, the participation rate in presidential elections has grown over time To illustrate, in the 1920s the average participation rate in presidential elections was about 47 percent, while in the 1960s the average turnout rate was around 62 percent. See Tullock (1967, p 108) for a more detailed discussion of the determinants of participation rates

[11] The participation rate is defined here as the number of votes cast divided by the voting-age population Since women were prohibited from voting prior to 1920, an adjustment had to be made to those rates to make them comparable with post-1920 rates The participation rate averaged 0 12 higher in 1924 and 1928 than in the 1916 election We therefore added 0 12 to pre-1920 rates in order to achieve reasonable comparability for the full period

[12] Voting is one way of expressing one's interest in an election; contributing to one's favorite candidate(s) is another To the extent that broadcast advertising generates campaign spending increases via this source, it is by no means clear why we (or Congress) should be concerned Presumably, enhanced voter awareness and participation (including contributing behavior) is a desirable outcome in a democracy

Vote difference.—The difference between the winning and the losing candidate's percentage share of the total vote—an ex post measure of the closeness of the race—also serves to control for nonbroadcasting influences on campaign expenditures[13] (U.S. Bureau of the Census, various issues). On the one hand, the closer potential voters-contributors perceive a race to be, *ceteris paribus*, the greater the likelihood they will see their votes and contributions as making a difference to the outcome. Consequently, closely fought contests may prove to be relatively expensive. On the other hand, campaigns in which one candidate appears to have a significant lead may (*a*) create a "bandwagon" effect that increases contributions to, and spending by, the leading candidate, and (*b*) encourage additional donations to, and spending by, the "underdog" in an effort to overtake the leader. Consequently, we have no prior expectations regarding the sign of this variable's coefficient.[14]

As with the participation-rate variable, there may be some reciprocal causation between the vote difference and campaign spending. An increase in (expected) vote difference could, for example, stimulate additional contributions to the leading candidate; however, additional donations to, and expenditures by, the leading candidate may, of course, widen the lead.

Other determinants.—In addition to the above-mentioned determinants, we experimented with a number of other variables that seemed on a priori grounds to be relevant but in practice had virtually no effect on either the explanatory power of the various estimating equations or the estimated coefficients of the other variables in the equations These variables, which we consequently omitted from the estimating equations, include (1) the level of opposition spending,[15] (2) a binary variable for incumbency, and (3) a binary variable indicating whether a "major" third-party candidate was running for the presidency.[16]

---

[13] This variable was constructed from the vote shares of the major parties only

[14] The ambiguity of the sign is implicit, e.g , in Plattner's statement (1974, p. 121) that "closer races tend to involve higher expenditures than more one-sided ones, a reversal of the causal relationship that is conventionally said to hold in this matter: *It is not that candidates win because they raise more money, so much as that they are able to raise more money because they are a good bet to win.*"

[15] The level of opposition spending would seem to have the potential for influencing the volume of contributions to any one candidate. Campaign spending, in part, serves the same function as many advertising expenditures, i e , maintenance of "market share " *Ceteris paribus*, an increase in contributions to a candidate might enhance his prospects (and those of his contributors) at the expense of his opponent. In turn, the contributors to the opponent may find it profitable to increase their contributions so that he can at least recover his initial market share. This duopoly situation could conceivably produce an escalation of campaign spending (contributions) with little or no effect on the candidates' respective market shares. Thus, a candidate's expenditures may well depend on the level of his opponent's spending. However, empirically this variable had virtually no effect on the regression results.

[16] "Major" third-party candidates ran in 1892 (Weaver), 1912 (Roosevelt), 1924 (La Follette), 1948 (Thurmond), and 1968 (Wallace) when the combined share of the

The price of radio and television advertisements relative to other campaign techniques is also a potentially important determinant not only of broadcast-related campaign expenditures but also of total campaign expenditures. Unfortunately, measures of this relative price are not available, so this variable was excluded from the estimating equations.[17]

B. The Findings

This section presents quantitative evidence that bears on the widely accepted proposition that the introduction of radio and then television occasioned substantial—even dramatic—increases in political campaign spending, perhaps most notably in presidential elections. In attempting to quantify the relationship between broadcast usage and campaign spending, we estimated a number of alternative models with data on presidential election campaigns from 1868 through 1972, excepting the 1944 war-year campaign as unrepresentative.

The regression results for several models in the linear form $Y = a_0 + a_1X_1 + \cdots + a_nX_n$ are reported in table 2.[18] A remarkable aspect of these findings is that the estimated coefficients of the broadcast usage variables are always negative and, with few exceptions, statistically significant. In striking contrast to the findings of simple correlation studies, these results—based on multivariate methods—suggest that the availability and usage of the broadcast media actually moderated the secular growth in presidential campaign spending! Perhaps over time, broadcast advertising—with its substantial "public good" characteristics—has tended to displace relatively more costly campaign techniques (e.g., direct mailings, personal appearances), thereby moderating the growth in campaign expenditures. The estimates reported in table 2 suggest that the attenuation effect of broadcast advertising on presidential campaign spending is considerable, approaching $25 million (1957

---

total vote going to the Democratic and the Republican candidates was 89.0, 65.1, 82.8, 94.7, and 86.1 percent, respectively (Congressional Quarterly Service, Inc. 1969, p. 122). The "third-party" binary, which took on a value of unity for these 5 years and zero otherwise, had virtually no effect on the regression results.

[17] See Backman (1967), e.g., for additional discussion regarding the limited availability of time-series data in this area.

[18] We also experimented with alternative functional forms of the estimating equations, including a model in which all variables (excepting the party-affiliation dummy) were expressed in logarithms and a model in which only total campaign spending—the dependent variable—was expressed in logarithmic form (to account for the probable underreporting of expenditures). The estimates provided by these alternative functional forms were qualitatively identical to those of the simple linear model, so we report in the text only the results for the simple linear specification.

TABLE 2

PRESIDENTIAL CAMPAIGN SPENDING EQUATIONS, 1868–1972

	MODEL			
	I	II	III	IV
Constant	−34.89	−35.00	−34.37	−29.11
TV use	−8.85 (1.74)	−10.95 (2.14)	−11.60 (2.56)	−13.11 (2.84)
Radio use	−17.08 (3.63)	−14.93 (3.18)	−14.27 (3.53)	−13.12 (3.17)
Participation rate	65.60 (3.53)	71.59 (3.80)	71.40 (3.83)	61.80 (3.31)
Vote difference	8.84 (1.26)	11.58 (1.64)	12.55 (2.05)	...
Federal spending	0.19 (1.90)		...	
Voting-age population	0.22 (1.73)	0.02 (0.29)		
Real GNP	−0.03 (0.81)	0.04 (2.81)	0.04 (4.40)	0.05 (4.72)
Party affiliation	3.41 (3.62)	3.41 (3.52)	3.41 (3.56)	3.41 (3.44)
R^2	.71	.69	.68	.66

NOTE.—Absolute value *t*-statistics in parentheses. Ellipses denote that variable did not enter that particular model. The dependent variable is reported national-level campaign spending on presidential general elections in millions of 1957 dollars, radio use = percentage of households with radios, TV use = percentage of households with televisions; participation rate = percentage of the voting-age population casting votes, vote difference = the difference in the percentage of the vote received by the winning and the losing candidates in the general election, federal spending = the federal government's outlays in billions of 1957 dollars, and party affiliation = 1 if a Republican candidate, 0 if Democratic.

dollars) per candidate in recent elections.[19] Notably, the attenuation estimates are rather insensitive to changes in model specification (see table 2), allowing us to dismiss the possibility that our findings arose merely from a fortuitous combination of explanatory variables.

If the broadcast media are not responsible for the recent growth in political campaign spending, what then is the source of this growth? Our findings offer some useful quantitative evidence with regard to this question also.

First, the coefficient of the participation rate is positive, as expected, and consistently significant. Since the participation rate (and, presumably underlying that, voter interest) in presidential elections has increased over the past several decades, some of the growth in campaign spending is no doubt traceable to this source.[20]

---

[19] If broadcast campaigning has affected any of the other explanatory variables, these attenuation estimates will be biased to some extent. E.g., political broadcast advertising may have contributed to a rise in the participation rate—a variable positively correlated with campaign spending. However, even if we assume that broadcasting is responsible for the entire increase in the participation rate between the 1920s and the 1970s, the attenuation estimates are reduced by only $9.3 million to $10.7 million per candidate, or by about 40 percent on average. Even under this extreme assumption, the net impact of broadcasting on presidential campaign spending remains strongly negative.

Second, the positive and significant coefficient of the federal spending variable—a proxy for availability of legislative profits—indicates that, as expected, growing governmental involvement in the economy stimulates campaign contributions and, thus, campaign spending. Thus, the rapid expansion in recent years in the federal government's role in the economy appears to underlie much of the growth in political campaign expenditures.

In addition, growth in real income and expansion in the size of the voting-age population have likely increased campaign giving and, therefore, spending. The regressions reveal that both of these variables (real GNP and voting-age population) generally have a significant positive influence on total campaign expenditures.

Finally, the positive coefficient of the vote difference variable suggests that one-sided contests tend, *ceteris paribus*, to be relatively costly. However, this observation must be tempered by the fact that vote difference is statistically significant in only one of three models.

Summary and Policy Implications

The multivariate analysis in this paper casts considerable doubt on the generally accepted hypothesis that a new campaign technique such as broadcast advertising, or the " 'push' of modernity," as one observer (Heidenheimer 1970, p 5) labels it, intensifies the level of political campaign spending Subject to the qualifications noted above,[21] our estimates support the alternative proposition, namely, that broadcast usage has mitigated the secular growth in campaign spending.

These findings suggest, for example, that the Federal Election Campaign Act of 1971, which limited expenditures on broadcast advertising, may well have yielded negligible—if not perverse—results. If broadcast usage has, in fact, moderated the growth in campaign spending, restrictions on radio and television advertising expenditures can be expected to heighten campaign spending. Nichols (1974, p. 148) has noted that "the limits in the Act did not *prevent* 1972 Presidential campaign expenditures from setting a new record" (emphasis added). According to our estimates, this "new record" is not surprising and may have been, in part, *caused* by the 1971 legislation. As a response to the Supreme Court ruling on the Campaign Act (as amended), a reimposition of the limitations on media expenditures by federal candidates is easily conceivable; however, in the interest of holding down the level of campaign spending, such a congressional reaction appears ill-advised in light of our findings.

---

[20] See n 10 above for an illustration of the extent to which participation rates have changed over time

[21] To reiterate some of the qualifications First, the study relies on reported national-level committee spending If the accuracy of reporting has varied over time due to changes in campaign spending laws and/or their enforcement, our results may be distorted Second, some of the independent variables may be structurally related. E.g., television and radio may have increased the participation rate, and, as a result, some part of the positive influence on campaign spending attributed to the participation rate may be due to the broadcast media Furthermore, increased competition for political office induced by radio and television usage might have stimulated candidates to generate additional legislative profits for the purpose of securing still greater contributions. While we recognize such shortcomings in the analysis, we nevertheless believe that it yields useful quantitative information about the effect of the broadcast media on political campaign spending

The 1974 amendments to the Federal Election Campaign Act attempted to limit total campaign spending itself. However, imposing a ceiling on total campaign expenditures seems analogous to "freezing" prices to stop an inflation The price freeze fails to remove the underlying causes of the inflation and, consequently, makes black-market activities attractively profitable. A ceiling on total campaign spending could very well occasion greater in-kind contributions of such things as labor services Moreover, it may also lead to increasing amounts of illegal "donations" by those seeking legislative profits—for instance, covertly given personal gifts, or promises of future favors (e g., jobs for relatives and friends) [22]

Some of our estimates suggest that increases in the availability of legislative revenues, generated by growing government involvement in economic affairs, are to some extent responsible for the growth in campaign contributions and expenditures. If so, the continuing expansion of governmental control over economic activity can be expected to further increase "donations" (whether legal or illegal) designed to capture legislative profits for the contributors

References

Alexander, H. E. *Political Financing* Minneapolis. Burgess, 1972

Backman, J. *Advertising and Competition* New York Univ London Press, 1967

Committee for Economic Development. *Financing a Better Election System*. New York· Committee Econ. Development, 1968.

Committee on Commerce, U.S. Senate. *Federal Election Campaign Act of 1973* (hearings before the Subcommittee on Communications). Washington. Committee on Commerce, 1973.

Congressional Quarterly Service, Inc. *Politics in America*. 3d ed. Washington. Congressional Q. Service, 1969.

Council of Economic Advisors. *Economic Report of the President*. Washington: Council Econ. Advisors, 1974.

Green, M. J., et al *Ralph Nader Congress Project: Who Runs Congress?* New York: Bantam, 1972.

---

[22] Stigler (1971, p 13) conjectures that much of the compensation to legislative leaders already takes the form of extrapolitical payments.

Heard, A. *The Costs of Democracy*. Chapel Hill: Univ. North Carolina Press, 1960.

Heidenheimer, A. J. "Major Modes of Raising, Spending and Controlling Political Funds during and between Election Campaigns." In *Comparative Political Finance*, edited by A. J. Heidenheimer. Lexington, Mass.: Heath, 1970.

Nichols, D. *Financing Elections*. New York: New Viewpoint, 1974.

Plattner, M. F "Campaign Financing—the Dilemmas of Reform." *Public Interest* 37 (Fall 1974). 112–30.

Scherer, F. M *Industrial Market Structure and Economic Performance* Chicago: Rand McNally, 1970.

Stigler, G. J "The Theory of Economic Regulation " *Bell J. Econ and Management Sci.* 2 (Spring 1971) 3–21.

Television Digest, Inc *Television Factbook* Washington Television Digest, various issues

Tullock, G. *Towards a Mathematics of Politics* Ann Arbor Univ Michigan Press, 1967.

U S. Bureau of the Census *Historical Statistics of the United States: Colonial Times to 1957* Washington. Bur. Census, 1960

——— *Long Term Economic Growth, 1860–1965.* Washington Bur. Census, 1966

——— *Statistical Abstract of the United States.* Washington Bur Census, various issues

Advertising, Concentration, and Price-Cost Margins

Allyn D Strickland and Leonard W Weiss

University of Wisconsin, Madison

The industrial organization literature contains many empirical tests of hypothesized relationships among elements of industry structure and performance Most of these studies used ordinary least-squares regression to estimate single-equation relationships This approach is incorrect if the relationship being estimated is part of a simultaneous-equations system This paper treats three important relationships in industrial organization as a single simultaneous-equations system. The results are consistent with those obtained from single-equation models and suggest that simultaneous-equations bias is not an important factor in the estimation of structure-performance relationships

The Model

The Need for a Simultaneous-Equations Model

Our original purpose was to estimate the concentration-advertising relationship In a previous article one of us (Weiss 1971) proposed that this relationship be examined on a simultaneous-equations basis The Dorfman-Steiner model (Dorfman and Steiner 1954) indicated that the optimal level of advertising expenditure for a firm would be where the marginal revenue product of advertising equaled the absolute value of the elasticity of demand for the firm's product Since the price elasticity of the demand curve faced by a firm is lower in more concentrated industries, advertising intensity might be expected to increase with concentration This effect might be reenforced to the extent that there are externalities in advertising That is, if advertising affects industry as well as firm demand, firms with large market shares would internalize a larger proportion of those industry effects

At the same time, there is likely to be feedback from advertising to concentration through the probable economies of scale in advertising—at least for heavily advertised national brands This suggests that a correct

[*Journal of Political Economy*, 1976, vol 84, no 5]
© 1976 by The University of Chicago All rights reserved

specification of the relationship would be a pair of simultaneous equations of the form

$$\frac{Ad}{S} = f\left(C, \frac{CD}{S}\right),$$
$$C = g\left(\frac{Ad}{S}, \frac{MES}{S}\right), \tag{1}$$

where Ad is advertising expense, S is value of shipments, C is concentration, CD is consumer demand, and MES is minimum efficient scale

A number of more recent analyses have also called for simultaneous-equations models Greer (1971) estimated a three-equation model where Ad/S, C, and growth were endogenous Three others (Cable 1972; Schmalensee 1972; Commanor and Wilson 1974) argue for a two-equation model where Ad/S and the price-cost margin are endogenous The argument is best expressed by Schmalensee, who concludes that optimal advertising intensity is

$$\frac{Ad}{S} = \frac{P - MC}{P}(a + \eta \bar{a}),$$

where P is price, MC is marginal production costs, a and $\bar{a}$ are elasticities of a firm's sales with respect to its own advertising and that of its rivals, respectively, and η is the elasticity of its rivals' advertising with respect to its own It follows that advertising will be more intense the higher the margin firms will advertise more the more that an additional unit sold adds to profits

Accepting this argument leads us to a three-equation model where Ad/S, C, and margin are endogenous, since there is ample theoretical basis for expecting both concentration and advertising to affect margins Our model will therefore take the form

$$\frac{Ad}{S} = f\left(C, M, \frac{CD}{S}\right),$$
$$C = f\left(\frac{Ad}{S}, \frac{MES}{S}\right), \tag{2}$$
$$M = f\left(C, \frac{Ad}{S}, X\right),$$

where M is the price-cost margin and X is one or more exogenous variables that affect M but not Ad/S or C. Good candidates for X are growth and capital intensity.

The Determinants of Advertising Intensity

Following Greer (1971) and Cable (1972), we expect the effect of concentration on advertising to take the form of an inverted U Advertising is expected to increase with concentration at first for the reasons given above but to decrease at very high levels of concentration because collusion to avoid mutually offsetting advertising becomes easier (Greer 1971) or merely because a pure monopolist would avoid all such mutually offsetting advertising (Cable 1972) This argument implies that the advertising equation should contain both C and C^2, with a positive sign expected for the coefficient of C and a negative sign for that of C^2

Following the arguments of Cable (1972), Schmalensee (1972), and Commanor and Wilson (1974), we also include price-cost margin (M) in the advertising equation, expecting its coefficient to bear a positive sign.

Finally, we include a number of variables intended to reflect advertising effectiveness The most important is surely the share of total sales going to consumers (CD/S) We expect it to have a positive effect because consumer goods appear to be more differentiable than producer goods and because advertising is probably the more effective way of reaching millions of consumers while salesmen may be a more effective way to reach the typically far fewer industrial buyers To control further for product differentiability, we will introduce growth (following Cable, Greer, and Commanor and Wilson) and a durability dummy (following Commanor and Wilson)

This leaves an advertising intensity equation of

$$\frac{Ad}{S} = a_0 + a_1 M + a_2 \frac{CD}{S} + a_3 C + a_4 C^2 + a_5 Gr + a_6 Dur, \qquad (3)$$

where Ad is advertising expense, S is value of shipments, CD is consumer demand, C is the four-firm concentration ratio, M is the price-cost margin (value added less payroll, divided by value of shipments), Gr is the average annual rate of growth in industrial production from 1954 to 1963, and Dur is a dummy variable equal to one for durable-goods industries and to zero for nondurable-goods industries We expect the signs of a_1, a_2, a_3, and a_5 to be positive and that of a_4 to be negative Following Commanor and Wilson, we give no a priori judgment about the sign of a_6

The Determinants of Concentration

Conventional price theory predicts that an important determinant of concentration will be optimal firm scale as a percentage of market size. If the long-run average cost curve is flat at sizes above the minimum efficient scale, then the four-firm concentration ratio may exceed four times minimum efficient scale. This implies that the correlation between concentration and estimated minimum efficient scale will not be perfect even in long-run equilibrium.

Since most of the economies of scale in production appear to be attained at the plant level, an estimate of plant minimum efficient scale will be used as a determinant of concentration. Direct estimates of minimum efficient scale are available for only a minority of industries, so some proxy is needed. The two most common proxies are the average plant size among the largest plants accounting for half of industry shipments or employment (Commanor and Wilson 1967), and the "midpoint plant size"—the size of the plant that is at the midpoint of the shipments size distribution. Half of total shipments come from plants larger than the midpoint plant size. Of the two, the midpoint plant size correlates more closely with available engineering estimates of minimum efficient scale.[1] We will therefore use that variable divided by industry shipments (*MES*/*S*) as our economies-of-scale proxy.

The original impetus for this study arose from the expectation that advertising intensity would increase concentration due to economies of scale in advertising. Such economies of scale are conceptually slippery, but Commanor and Wilson (1974, chap. 10) provide evidence suggesting that they may be large in industries where advertising intensity is high.

In addition to probable economies of scale in advertising, strong product differentiation in and of itself might well be associated with high concentration. Firms that are successful in differentiating their products are also apt to be successful in attaining large market shares. There appears to have been a consistent postwar trend toward increasing concentration in highly differentiated products, although average concentration in manufacturing as a whole hardly changed during this period (Mueller and Hamm 1974). This trend is most likely attributable to the increase in advertising intensity that came with the television revolution. The effect of advertising on concentration is only likely to be important in industries where advertising intensity is great. In sum, there seems to be ample justification for including *Ad*/*S* as an independent explanatory variable in the concentration equation.

---

[1] In an as yet unpublished paper, one of us (Weiss 1975) correlates the two measures with direct estimates of *MES* made by Pratten (1971), Scherer (1973), and unpublished estimates of his own. The R^2's relating the two measures with the three sets of direct estimates are given below.

	Scherer	Pratten	Weiss
Number of industries	12	19	16
Midpoint plant size	.9033	.4338	.4414
Average size of plants in top half of shipments	.7671	.3808	3974

Altogether, then, our concentration equation will be

$$C = b_0 + b_1 \frac{Ad}{S} + b_2 \frac{MES}{S} \tag{4}$$

We expect both b_1 and b_2 to have positive coefficients

The Determinants of Price-Cost Margins

The literature contains a large number of statistical studies in which profit rates or price-cost margins are explained by concentration, advertising, and certain other variables (see Weiss 1974). Our margin variable will be the widely used "price-cost margin" based on census totals. It can be precisely measured at the four-digit level, and it avoids most of the accounting difficulties of corporate profit margins Since the cost of capital is included in this margin, an essential independent variable is a measure of capital intensity, K/S In our study, this variable is the gross fixed value of assets divided by value of shipments Similarly, output growth due either to unanticipated increases in demand or to unanticipated decreases in costs can be expected to result in high margins. Our measure of growth is average annual growth rates in output from 1954 to 1963 based on the census index of industrial production

There are strong theoretical grounds for believing that market structure will affect price-cost margins Concentration is probably the most important element of market structure in this regard Although there are many oligopoly theories, virtually all of them predict an increase in the effectiveness of collusion (a decrease in the cost of collusion) as concentration rises The area of disagreement concerns the implied functional form of the concentration-margin relationship rather than the direction of its effect

The concentration ratios used refer to national industries. These clearly do not characterize market concentration correctly where markets are local or regional in character To allow for this, we follow Collins and Preston (1969) in introducing a measure of the geographic dispersion of output (GD) as well

Finally, the attainable margins of effectively colluding firms can be expected to rise with the barriers to entry. The classic barriers to entry are those associated with product differentiation and scale requirements As in many other studies, advertising intensity will serve as our proxy for the product differentiation barrier Since advertising is included in the margin, its coefficient should equal 1 0 even if there is no product differentiation effect. It should significantly exceed 1 0 if the product differentiation barrier has its expected effect. The scale barrier is MES/S once more

Altogether, then, our price-cost margin equation will be

$$M = c_0 + c_1 \frac{K}{S} + c_2 Gr + c_3 C + c_4 GD + c_5 \frac{Ad}{S} + c_6 \frac{MES}{S}. \quad (5)$$

We expect all of the coefficients to have positive signs, except for c_4, which should be negative.

The Structure and Estimation of the Simultaneous-Equations Model

In our three structural equations, the price-cost margin appears as a determinant of advertising intensity, both concentration and advertising intensity appear as determinants of price-cost margins, and advertising intensity is one of the determinants of concentration. We feel that there is good theoretical justification for inclusion of these variables as shown. Since the same variables also appear as dependent variables in our three equations, we must treat advertising intensity, concentration, and price-cost margins as endogenous variables whose values are *jointly* determined in a simultaneous-equations system consisting of equations (3), (4), and (5) The order conditions for identifiability indicate that equation (4) is overidentified while equations (3) and (5) are exactly identified

Our three-equation system is linear in parameters but nonlinear in endogenous variables because of the squared concentration term in equation (3) As a result, the reduced-form equations are functions of the square root of linear relations among the endogenous variables. The reduced-form equations can, however, be approximated by a polynomial function of the exogenous variables (Kelejian 1971) This will lead to consistent estimates of the system's parameters using two-stage least-squares estimation We used second-degree polynomials to approximate the reduced-form equations.

Empirical Analysis

The Data

All of the variables used in these equations are derived from the 1963 Census of Manufactures, except for *Ad/S*, *CD/S*, and *Dur*, the durability dummy. Values for *Ad/S* and *CD/S* are taken from the 1963 input-output tables for "detailed industries." *Ad/S* is just the direct requirements coefficient of each manufacturing sector for inputs from the sector labeled "advertising." *CD/S* is "personal consumption expenditures" divided by "total output," where figures for both are taken from the transactions matrix of those tables. The durability dummy was assigned on the basis of judgment.

All of the census variables are for four-digit industries. Input-output sectors correspond exactly to four-digit industries in 230 cases. In the remaining cases,[2] sectors and census industries do not match perfectly. In all but one case, two or more census industries are assigned to a single sector. As a result, our advertising and consumer demand variables for industries that are combined in the input-output sectors are more aggregated than the other variables

The year 1963 was chosen because all the variables used in this study are available for that year and because most of the conventional relationships hold up well when estimated as single-equation models in that year It was a year of moderate prosperity with little inflation—a decade after the last bout of serious inflation and price controls—so price-cost margins had the opportunity to adjust toward equilibrium levels Similarly, the television revolution in advertising was more than a decade old, so that Ad/S had probably come close to equilibrium levels

The sample consists of 408 of the 417 four-digit Standard Industrial Classification (SIC) manufacturing industries from the 1963 Census of Manufactures Of the nine industries deleted from the sample, two (2819, 3943) were dropped for lack of advertising data The seven remaining industries (2814, 3332, 3334, 3492, 3636, 3723, 3942) could not be used, since the census's disclosure requirement prevented the construction of four firm concentration ratios

OLS Results

Table 1 shows ordinary least-squares estimates of the three structural equations Most of the coefficients of conventional variables in the OLS regressions bear the expected signs and are significantly different from zero Advertising rises with margins, as predicted by Cable, Schmalensee, and Commanor and Wilson And, as expected by Cable and Greer, the relation between concentration and advertising intensity is of an inverted-U form Advertising intensity reaches its peak at $C = 0.49$ As expected, the variable CD/S has a very strong effect on advertising. Advertising also increases with growth, though the effect is weak Durability has a nonsignificant negative effect

The results for equation (4) are also as expected Both advertising intensity and MES/S have significant positive effects on concentration.

In equation (5), margins rise with both concentration and advertising intensity The advertising coefficient exceeds 1.0 by a significant amount (by 4.35 standard errors), suggesting that there is a substantial product differentiation barrier to entry associated with advertising. Growth, MES/S, capital intensity, and geographic dispersion all have the expected

---

[2] Industry 2819 (inorganic chemicals, not elsewhere classified) was excluded because it was split into two sectors in the input-output tables Industry 3934 (baby carriages) was excluded because it does not appear in the input-output table for some reason. Except for these two cases, our sample corresponds exactly to that of Collins and Preston (1969)

TABLE 1

OLS ESTIMATES OF OUR THREE EQUATIONS (*t*-RATIOS IN PARENTHESES)

	DEPENDENT VARIABLE		
	Ad/S Eq (3)	*C* Eq. (4)	*M* Eq (5)
Constant	−0.0314 (−7 45)	0 2698 (25 93)	0 1682 (17 15)
C	0 0554 (3 56)		0 0629 (2 89)
C^2	−0 0568 (−3 38)		
M	0 1123 (9 84)		
CD/S	0 0257 (8 94)		
Gr	0 0387 (1 64)		0 2255 (2 61)
Dur	−0 0021 (−1 11)		
Ad/S		1 1613 (3 33)	1 6536 (11 00)
MES/S		4 1852 (18 99)	0 0686 (0 54)
K/S			0 1123 (8 03)
GD			−0 0003 (−2 90)
R^2	374	485	402
df	401	405	401

effects, though the coefficient of *MES/S* could easily differ this much from zero by chance

Two-Stage Least-Squares Results

The results in table 1 may be biased because of the simultaneous character of our model Two-stage least-squares estimates which avoid this bias appear in table 2

The effect of concentration on advertising found in the OLS estimates remains in the 2SLS estimate of equation (3), but advertising now reaches its maximum at a concentration ratio of 0 57 The effect of margins on advertising is greatly reduced, though it is still significant The effects of *CD/S*, growth, and durability are little altered

Equation (4) is little changed in the two-stage least-squares estimates The effect of advertising on concentration is somewhat enhanced and remains statistically significant[3]

---

[3] At the suggestion of a referee, we also tried a different form of the concentration equation (eq. [4]) where *CD/S* was included among the independent variables All three independent variables were highly significant in both the OLS and the 2SLQ regressions The coefficient of advertising had a positive sign and that of *CD/S* had a negative sign We do not know how to interpret this result

TABLE 2

TWO-STAGE LEAST-SQUARES ESTIMATES OF OUR THREE EQUATIONS
(t-RATIOS IN PARENTHESES)

	DEPENDENT VARIABLE		
	Ad/S Eq. (3)	*C* Eq. (4)	*M* Eq. (5)
Constant	−0.0245	0.2591	0.1736
	(−3.86)	(21.30)	(14.66)
C	0.0737		0.0377
	(2.84)		(0.93)
C^2	−0.0643		
	(−2.64)		
M	0.0544		
	(2.01)		
CD/S	0.0269		
	(8.96)		
Gr	0.0539		0.2336
	(2.09)		(2.61)
Dur	−0.0018		
	(−0.93)		
Ad/S		1.5347	1.6256
		(2.42)	(5.52)
MES/S		4.169	0.1720
		(18.84)	(0.92)
K/S			0.1165
			(7.30)
GD			−0.0003
			(−2.79)

The last equation does change in an important way, however. The effect of concentration on the price-cost margin remains positive but could easily be due to chance. In addition, the coefficient of advertising intensity now exceeds 1.0 by an amount that is barely significant using a one-tailed test (by 1.91 standard errors), thus raising some question about the product differentiation barrier. The coefficients of growth, *MES/S*, *K/S*, and geographic dispersion remain close to their values and significance levels found in the OLS regressions.

The reason for the weak relation between concentration and margins in the third equation is probably collinearity. Observed concentration and the other independent variables in the OLS estimate of equation (5) are related by an R^2 of .532, but this rises to .794 when $\hat{C}$ is regressed on $\widehat{Ad/S}$ and the other independent variables in the second-stage least-squares estimate of equation (5). This is because the estimated value of $\hat{C}$ is primarily determined by *MES/S* and, to a lesser extent, *K/S* and *GD*. In effect, equation (5) is not fully identified, since the exogenous variables excluded from equation (4) are not importantly correlated with concentration. We conclude that the effect of concentration on margins cannot be distinguished from the effect of optimal plant scale.

The same does not explain the weakened effect of advertising on margins. It is related to the other independent variables in the OLS estimate of equation (5) by an R^2 of only .06639 and to $\hat{C}$ and the other independent variables in the two-stage least-squares estimate by an R^2 of only .18118.

Consumer and Producer Goods Separately

The arguments for our model seem most relevant to consumer goods and much less so for producer goods Specifically, one would not expect a strong effect of concentration on advertising, of advertising on concentration, or of advertising on margins in producer-goods industries where advertising is normally minor In addition, the effect of concentration on margins has generally been found to be much greater in consumer- than in producer-goods industries (Collins and Preston 1969; Weiss 1974), presumably because of the difficulty of collusion where buyers are large and well informed

For the purposes of this study, consumer-goods industries are defined as those where $CD/S \geq 0.5$ Breaking the sample in this way leaves 102 consumer-goods industries and 306 producer-goods industries

Table 3 shows OLS and two-stage least-squares estimates for consumer goods Table 4 shows the same for the producer-goods industries.

The consumer-goods regressions are similar to those for the overall sample except for the puzzling negative sign on CD/S in the advertising equation We presume that this anomalous result is due to a lack of variation in the sample where $CD/S \geq 0\ 50$

The effect of concentration on advertising is much stronger in the consumer-goods sample, but the coefficients retain their expected signs in the producer-goods sample. Advertising intensity reaches maximum at $C = 0\ 37$ in the OLS estimates and $C = 0\ 46$ in the 2SLS estimates for consumer goods and at concentration ratios of 0.47 and 0 72 for producer goods Margins are also much less important in determining advertising in the consumer-goods sample On the other hand, the effect of growth on advertising is significant in only the producer-goods sample, suggesting that producer-goods advertising is more concentrated on new products and new producers than in consumer-goods industries

As expected, advertising has a greater effect on concentration in consumer-goods industries, but it remains significant in the producer goods sample.

The effect of concentration on margin is significant in the OLS estimate and nonsignificant in the 2SLS estimates in both samples The explanation suggested above seems applicable here also, since the R^2s relating to Ad/S and the other independent variables are .799 in the consumer goods sample and 779 in the producer-goods sample[4]

The coefficient of advertising in the margin equation significant exceeds 1.0 in the OLS estimates but is not significantly greater than 1

TABLE 3

ESTIMATES FOR 102 CONSUMER-GOODS INDUSTRIES
(t-RATIOS IN PARENTHESES)

	DEPENDENT VARIABLE					
	OLS Estimates			2SLS Estimates		
	Ad/S Eq (3)	*C* Eq (4)	*M* Eq (5)	*Ad/S* Eq (3)	*C* Eq. (4)	*M* Eq (5)
Constant	−0.0005	0.2165	0.1704	−0.0019	0.1748	0.1783
	(−0.02)	(12.13)	(8.66)	(−0.08)	(6.77)	(8.45)
C	0.1446		0.1376	0.1761		0.0953
	(3.29)		(3.00)	(3.18)		(1.43)
C^2	−0.1972			−0.1918		
	(−3.98)			(−3.62)		
M	0.3159			0.2146		
	(10.51)			(3.34)		
CD/S	−0.0864			−0.0703		
	(−3.45)			(−2.57)		
Gr	0.0562		−0.0077	0.0843		0.0385
	(0.97)		(−0.51)	(1.36)		(0.23)
Dur	−0.0134			−0.0152		
	(−2.44)			(−2.65)		
Ad/S		1.6740	1.6151		3.3710	1.3959
		(4.12)	(8.73)		(4.18)	(2.90)
MES/S		3.7373	0.3726		3.5922	0.5203
		(11.17)	(1.67)		(9.76)	(1.89)
K/S			−0.0466			−0.0042
			(−0.83)			(−0.06)
GD			−0.0002			−0.0002
			(−1.23)			(−1.35)
R^2	.631	.611	.654			
df	95	99	95			

in the 2SLS estimate for the consumer-goods sample and is barely so by a one-tailed test in the producer-goods sample. Thus, more doubt is cast on the importance of the product differentiation barrier.

Surprisingly, capital intensity and geographic dispersion affect margins significantly in only the producer-goods industries.

Conclusions

It would appear that advertising increases with concentration over the range of *C* in which most observations fall and that advertising leads to greater concentration. The failure of earlier studies (Telser 1964) to obtain this result is presumably due to the crudeness of their data on advertising (roughly three digits in coverage), their failure to use a quadratic form, and their use of single-equation models. The effect of advertising on concentration suggests that the economies of scale in advertising are substantial.

On the other hand, the effect of advertising on margins is only mildly greater than what might be expected from the fact that advertising

TABLE 4

ESTIMATES FOR 306 PRODUCER-GOODS INDUSTRIES
(t-RATIOS IN PARENTHESES)

	DEPENDENT VARIABLE					
	OLS Estimates			2SLS Estimates		
	Ad/S Eq. (3)	*C* Eq. (4)	*M* Eq. (5)	*Ad/S* Eq. (3)	*C* Eq. (4)	*M* Eq. (5)
Constant	−0.0062 (−2.20)	0.2707 (21.30)	0.1715 (15.11)	−0.0001 (−0.04)	0.2661 (19.02)	0.1723 (13.16)
C	0.0113 (1.14)		0.0572 (2.31)	0.0036 (0.23)		0.0601 (1.33)
C^2	−0.0121 (−1.159)			−0.0025 (−0.17)		
M	0.0337 (4.38)			0.0110 (0.77)		
CD/S	0.0538 (11.49)			0.0563 (11.40)		
Gr	0.0361 (2.32)		0.2661 (2.60)	0.0446 (2.71)		0.2753 (2.67)
Dur	−0.0008 (−0.70)			−0.0006 (−0.53)		
Ad/S		1.5442 (2.15)	1.9540 (6.45)		2.0644 (2.09)	1.7782 (4.15)
MES/S		4.3659 (15.91)	−0.1334 (−0.87)		4.3546 (15.88)	−0.1423 (−0.65)
K/S			0.1201 (7.82)			0.1188 (7.24)
GD			−0.0003 (−2.87)			−0.0003 (−2.87)
R^2	408	464	330			
df	299	303	299			

expense is included in the price-cost margin and is not significantly so in consumer-goods industries. This suggests that the product differentiation barrier to entry is not very great.

Our finding with respect to the impact of concentration on margins seems equivocal. It cannot be distinguished from the effect of the plant scale variable in our second-stage regressions. The combined effect of *C* and *MES/S* is unequivocally positive and significant. The *F*-statistic for the two variables together is 36.14 with 2 and 401 degrees of freedom—easily significant at the .01 level. We have no basis for assigning their joint effect between concentration and the scale barrier. We presume that both are present.

References

Cable, J. "Market Structure, Advertising Policy, and Inter-Market Differences in Advertising Intensity." In *Market Structure and Corporate Behavior*, edited by K. Cowling. London: Grays Mills, 1972.

Collins, N. R., and Preston, L. E. "Price Cost Margins and Industry Structure." *Rev. Econ. and Statis.* 51 (August 1969): 271–86.

Commanor, W. S., and Wilson, T. A. "Advertising, Market Structure and Performance." *Rev. Econ. and Statis.* 49 (November 1967): 423–44.

———. *Advertising and Market Power.* Cambridge, Mass.. Harvard Univ Press, 1974.

Dorfman, R., and Steiner, P. O. "Optimal Advertising and Optimal Quality." *A.E.R.* 44 (December 1954) · 826–36

Greer, D. "Advertising and Market Concentration" *Southern Econ J* 38 (July 1971): 19–32.

Kelejian, H H. "Two Stage Least Squares and Econometric Systems Linear in Parameters but Non-linear in the Endogenous Variables" *J American Statis. Assoc.* 69 (June 1971) · 373–74

Mueller, W F., and Hamm, L. G "Trends in Industrial Market Concentration, 1947–1970." *Rev Econ and Statis.* 56 (November 1974). 511–20.

Pratten, C. F. *Economies of Scale in Manufacturing Industry* Cambridge. Cambridge Univ. Press, 1971.

Scherer, F M "The Determinants of Industrial Plant Sizes in Six Nations." *Rev. Econ. and Statis* 55 (May 1973) 135–45

Schmalensee, R *The Economics of Advertising* Amsterdam · North-Holland, 1972.

Telser, L "Advertising and Competition." *J.P E* 72 (December 1964) 537–62

Weiss, L W "Quantitative Studies in Industrial Organization." In *Frontiers of Quantitative Economics*, edited by M. Intriligator Amsterdam North-Holland, 1971

——— "The Concentration-Profits Relationship and Anti-Trust." In *Industrial Concentration · The New Learning*, edited by H. J. Goldschmid, H. M. Mann, and J F Weston. Boston · Little, Brown, 1974

———. "Optimal Plant Size and the Extent of Sub-optimal Capacity" Unpublished, 1975

Market Organization and the Effect of Import Quotas: Empirical Evidence

Susumu Hondai

Michigan State University

Standard economic theory states that the output of a final product and the purchase of a production factor are reduced (1) if competition in a final-product market is replaced by a monopoly, (2) if competition in the purchase of a production factor is replaced by a monopsony, or (3) if competition in the final-product and the factor markets is replaced simultaneously by a monopoly and a monopsony (Ophir 1969). This theory also states that, if (1) the elasticity of the foreign supply of a final product is large enough and (2) its import price is lower than the domestic monopolist price but higher than the competitive market price, the monopolist output will be greater, and the final-product price and the monopoly profits will be lower, with imports than without imports. On the other hand, the price of a production factor rises as a monopolist increases the output of a final product. So the relaxation of an import restriction leads to an increased factor price, which will then raise the income of the factor producers. However, if both the factor and the final-product markets are competitive, the relaxation of an import restriction leads to a lower income for factor producers. The effect of an import restriction on the income of factor producers thus depends on the structure of the factor and the final-product markets.

In this paper, I will contrast the impact of import quotas on the Japanese dairy product market from 1952 to 1971 and on the Japanese beef market from 1955 to 1971. The import quotas on dairy and beef products were originally designed to raise the income of the factor producers, the farmers who supplied their products to the dairy distributing and processing industry and the meat packing and processing industry.

Most agricultural products have two demand curves: one for fresh products and one for processed products. In figure 1, $D_0D'_1D_1$ is the fresh-and-processed aggregated demand curve for an agricultural product. Section $D_0D'_1$ of $D_0D'_1D_1$ is the fresh-product demand curve, while

I would like to thank William Haley and Harry G. Johnson for helpful comments on previous drafts of this paper.

[Journal of Political Economy, 1976, vol. 84, no. 5]
© 1976 by The University of Chicago. All rights reserved.

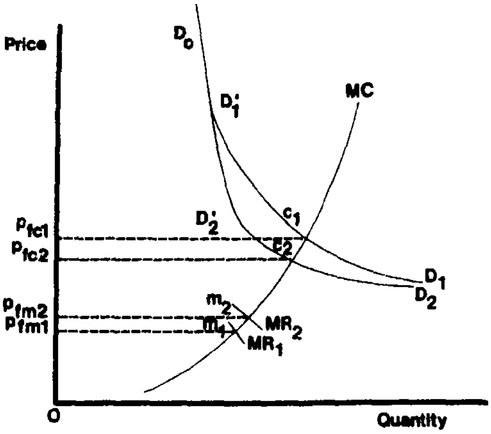


Fig 1

section D'_1D_1 is the fresh-and-processed-product demand curve In general, the demand elasticity for a fresh product is smaller than that for a processed product The marginal cost curve of a production factor for a distributing and processing industry is *MC* If a distributing and processing industry is competitive, its marginal revenue is equivalent to its demand curve Since the equilibrium point without imports is at c_1, the industry pays p_{fc1} to factor producers A small import quota of processed products shifts the demand curve to $D_0D'_2D_2$ and the equilibrium point to c_2 The factor price will fall from p_{fc1} to p_{fc2} Thus the quota will reduce the income of factor producers, the farmers

If the distributing and processing industry forms a monopoly, its marginal revenue equals its marginal cost at the equilibrium point Since the equilibrium point without imports is m_1, the industry pays p_{fm1} to factor producers The introduction of an import quota shifts the demand curve from $D_0D'_1D_1$ to $D_0D'_2D_2$ and the marginal-revenue curve from MR_1 to MR_2, if the demand curve is concave from above and if the import is a relatively small part of domestic consumption (Finger 1971) As a result, the factor price will rise from p_{fm1} to p_{fm2}, and the income of factor producers will also rise So the relaxation of an import restriction increases the income of the factor producers

The three biggest milk distributors and processors in Japan have a huge share of the milk product market They hold about 60, 95, 80, and 75 percent of the shares in the fluid whole milk, dried milk, butter, and cheese markets, respectively In the factor market, they purchase over 60 percent of the farm milk. The price differences between domestic and foreign dairy products, as shown in table 1, reflect partly import policies

TABLE 1

RATIOS OF DOMESTIC WHOLESALE PRICES TO IMPORT PRICES FOR DAIRY PRODUCTS AND BEEF

Product	1967–70 Average Price Ratio	1967–70 Average Market Share* (%)
Butter	2.48	79.1
Cheese	1.23	73.1
Dry milk	4.33	96.5
Beef	1.93	Competitive†

SOURCE.—Japan Ministry of Agriculture and Forestry 1971, Japan Ministry of Finance 1968–71.

* By the three largest distributing and processing firms.

† *Data on market shares for the beef industry are not available, since no firm has any significant share in the industry.*

and partly the market structure of the industry. In contrast, the beef market is competitive, so the price difference between domestic and foreign products is due mainly to the beef import-quota system. Because only natural cheese is protected by a tariff, the smallest difference is in the natural and the processed cheese prices. For dried milk, the three major firms share over 95 percent of the market, and the ratio of domestic price to import price exceeds 400 percent.

If an industry forms a monopsony in its factor markets or a monopoly in its final-product markets, the market model consists of the following equations:

$$MR = P_d\left(\frac{b_1 - 1}{b_1}\right) = f(Q_d, X_1), \tag{1}$$

$$MC = P_s\left(\frac{e_1 + 1}{e_1}\right) = g(Q_s, X_2), \tag{2}$$

$$MR = MC \text{ or } Q_d = Q_s, \tag{3}$$

where MR = industry marginal revenue, MC = factor marginal cost of the industry, P_d = final-product price, P_s = factor price; Q_s, Q_d = quantities supplied and demanded, respectively, b_1 = price elasticity of demand for a final product; e_1 = price elasticity of a factor supply; and X_1, X_2 = sets of exogenous variables which shift the MR and the MC curves, respectively. For the dairy markets, variable X_1 includes consumer income (Y), the price of substitute (P_{sub}), and import quota (Q); whereas variable X_2 includes the number of cows (NC), labor cost (IWR), and feed price (FC) (Japan Bureau of Statistics 1955–72; Japan Ministry of Agriculture and Forestry 1952–73). From equations (1), (2), and (3), we can derive reduced-form equations for the factor and the final-product prices. Since we are interested in the sign of an import quota, we divide both sides of the reduced-form equations by either $e_1/(e_1 + 1)$ or $b_1/(b_1 - 1)$ to get the following equations:

$$P_s = h_1(Y, Q, P_{sub}, IWR, NC, FC), \tag{4}$$

$$P_a = h_2(Y, Q, P_{sub}, IWR, NC, FC) \tag{5}$$

Using the linear form of equations (4) and (5), the impact of import quotas on farm and consumer milk prices can be estimated The results of the ordinary least-squares (OLS) estimations are listed in table 2.[1] The signs of the number of cows, the feed price index, and consumer income correspond to those implied by standard economic theory. The signs of the import quota are positive, indicating that the factor price rises as the import quota expands This indicates that the marginal-revenue curve of the industry shifts from MR_1 to MR_2 in figure 1 On the other hand, the number of cows and consumer income significantly affect the consumer price [2] Even though the signs of the import quota are negative, the expanded import quota does not significantly influence the consumer price

For the beef market, I used the number of beef cattle of the previous year (NB_{-1}), the number of milk cows of the previous year (NC_{-1}), import quotas of the current year and the previous year (Q and Q_{-1}), feed price index (FC), consumer income (Y), and consumer pork price (P_{sub}) (Japan Bureau of Statistics 1955–72, Japan Ministry of Agriculture and Forestry 1952–73) [3] Since slaughtered cows enter the beef market, an increased number of cows will shift the beef supply curve to the right But dairy cows utilize the same inputs as beef cattle, so an increased number of cows will pull resources from cattle production and raise resource prices Because the number of cows has increased rapidly since 1950, the latter should have a greater effect than the former. So the price of beef cattle should go up as the number of cows increases, and the sign of the number of cows is positive

---

[1] Variables used as independent variables are import quotas, number of cows, feed price index, and consumer income for the farm price model, import quotas, number of cows, and consumer income for the consumer price model Quantities and income variables are divided by population Price and income variables are deflated by the Japanese consumer price index (1970 = 1 00) Consumer income is the per capita income, before tax, of nonfarm families (Japan Bureau of Statistics 1955–72) I feel that this figure represents consumer disposable income in Japan better than other income series such as the gross domestic product, because there is no consistent gross domestic product figure available from 1951 to 1971

[2] The consumer price of milk products is the weighted average price which consumers paid for whole fluid milk, butter, skimmed milk, whole dried milk, and cheese. The prices of the latter four products are equivalent to fluid milk prices. The ratios of fluid milk for various uses to the total fluid milk generate the weights for prices.

[3] As in the previous case, quantities and income variables are divided by population The price and income variables are deflated by the Japanese consumer price index (1970 = 1.00). The linear form is also used in this case

TABLE 2

COEFFICIENTS OF MILK PRICES RESULTS OF OLS ESTIMATION

Dependent Variable	Constant	Import Quota (Q or Q_{-1})	N Cows (NC)	Feed Price (FC)	Income (Y)	$\bar{R}^2$	D-W
Farm price (P_s)	386.48	12.64*	−47.47	25.35	2.45	.73	1.99
		(3.24)	(8.67)	(111.16)	(0.35)		
	420.85	7.12†	−35.65	22.07	1.85	.56	1.72
		(3.94)	(10.18)	(143.88)	(0.39)		
Consumer price (P_c)	146.07	−0.390*	−4.68		0.110	.95	1.99
		(0.454)	(0.931)		(0.048)		
	145.20	−0.273†	−5.01		0.128	.95	2.01
		(0.431)	(0.771)		(0.042)		

NOTE.—Figures in parentheses indicate standard errors. Mean values of variables farm price = 498.01 yen/10 kg, consumer price = 119.29 yen/kg, import quota = 4.57 kg/person, N cows = 10.11/1,000 persons, feed price index = 1.39 (1970 = 1.00), and consumer per capita income = 202,740 yen

* The import quota of the current year

† The import quota of the previous year

TABLE 3

Coefficients of Beef Prices Results of OLS Estimation

Dependent Variable	Constant	Import Quota (Q or Q_{-1})	N Beef Cattle (NB_{-1})	N Cows (NC_{-1})	Feed Price Index (FC)	Income (Y)	Consumer Pork Price (P_{sub})	$\bar{R}^2$	D-W
Farm price (P_s)	326 13	−4 38* (57 67)	−18 84 (2 95)	13 24 (8 07)	265 03 (176 38)	−0 027 (0 312)	−0 275 (0 891)	96	1 83
	198 34	−135 93† (67 51)	−19 71 (2 49)	18 62 (6 64)	369 26 (144 72)	0 088 (0 229)	−1 004 (0 747)	.97	2 24
Consumer price (P_a)	1033 19	3 22* (120 93)	−34 38 (6 18)	3 98 (16 91)	229 33 (369 85)	0 843 (0 653)	2 529 (1 87)	97	1 29
	1034 23	8 95† (167 74)	−34 29 (6 18)	3 82 (16 50)	226 36 (359 62)	0 844 (0 570)	2.554 (1 855)	97	1 30

Note —Figures in parentheses indicate standard errors Mean values of variables farm price = 355 36 yen/kg, consumer price = 978 85 yen/kg, import quota = .136 kg/person, N cattle = 23 05/1,000 persons, V cows = 10 44/1,000 persons, feed price index = 1 34 (1970 = 1 0), consumer per capita income = 218,310 yen, and consumer pork price = 81 37 yen/100g

* The import quota of the current year

† The import quota of the previous year

TABLE 4

COEFFICIENTS OF MILK AND BEEF PRICES RESULTS OF THE TWO-STAGE LEAST-SQUARES ESTIMATION

Dependent Variable	Constant	Milk or Beef Import Quota (Q)	Consumer Milk or Beef Price (P_c)	Income (Y)	Feed Price Index (FC)	N Cows (NC)	N Beef Cattle (NBC_{-1})	N Cows (NC_{-1})	Consumer Pork Price (P_p)	Previous Year's Milk or Beef Import Quota (Q_{-1})	R^2	D-W
Milk import quota (Q)	42.05		−0.278 (0.152)			−0.568 (0.452)				0.337 (0.237)	.68	2.02
Farm milk price (P_s)	343.12	17.69 (4.84)		2.73 (0.42)	41.95 (120.33)	−53.34 (10.13)					.75	2.26
Beef import quota (Q)	0.414		−0.0003 (0.0009)	0.0022 (0.0011)		0.0116 (0.0320)				0.719 (0.362)	.66	2.10
Farm beef price (P_s)	220.29	−46.40 (69.69)		0.097 (0.338)	320.94 (187.51)		−18.47 (3.04)	15.99 (8.62)	−0.611 (0.960)		.97	2.10

NOTE.—Mean values of variables are the same as in tables 2 and 3, and the linear form is used for the estimation above. Figures in parentheses indicate asymptotic standard errors. Excluded exogenous variables are industrial wage rate and the farm beef price for the milk model, and industrial wage rate, farm milk price, and per capita fish harvested for the beef model. Coefficients of import-quota equations may be inefficient estimates due to the presence of the lagged dependent variables.

In table 3, the results show that the sign on the import quota is negative for the farm price and positive for the consumer price. But only the coefficient for the previous year's quota in the farm price equation differs from zero Thus the quota is likely to reduce the farm price of beef cattle

It is sometimes argued that the import quota is endogenous, since the government may expand the import quota as domestic consumer prices for the products go up So we will treat the import quota on milk and beef products as an endogenous variable and use the two-stage least-squares estimation method [4] In table 4, the results of the estimation show that the signs for the consumer milk price and the consumer beef price are negative for the milk and beef import quotas, respectively, but neither coefficient differs significantly from zero On the other hand, the import-quota sign is positive for the farm milk price and negative, but not significantly different from zero, for the farm beef price. Therefore, as the import quota expands, the price paid to farmers for milk increases

Two different assumptions, (1) the import quota is an exogenous variable and (2) the import quota is an endogenous variable, produce the same relationship among the import quota, the final-product price, and the factor price Despite government intentions, the import quota on dairy products does not increase the income of the dairy farmers Importing a final product introduces foreign competition for the domestic distributors and processors and decreases their income; however, the effects of importing a final product on the income of the factor producers depends on market structures in both factor and final-product markets

Contrary to common speculation, the import quota on milk products does not protect Japanese dairy farmers Both farmers and consumers, in this case, would be better off if the import quota were replaced with a tariff By keeping the tariff low, the Japanese government could impose severe limitations on the oligopolistic power of the processors while still protecting the domestic farmers Presently, if the quota on beef is maintained, some processors may grow large enough to control the market as beef consumption expands To prevent exploitation by big processors, the government can replace the quota with a tariff and thus create a more competitive market

---

[4] Import quota and farm price equations were defined as parts of a supply-demand model The equations of the milk-market model are as follows

$$Q = f_3(P_a, Q_{-1}, NC),$$

$$P_f = f_4(Q, NC, Y, FC),$$

where Q_{-1} is the import quota of the previous year, and other notations are the same as in the previous case. Then I estimated the equations above, which have consumer milk price (P_a), import quotas (Q), and farmer milk price (P_f) as endogenous variables, using the two-stage least-squares method The equations for the beef market model are the same, except they use the variables listed in table 4

References

Finger, J. M. "Protection and Domestic Output" *J. Internat Econ.* 1 (August 1971). 345–51.

Japan Bureau of Statistics, Office of the Prime Minister *Annual Report on the Family Income and Expenditure Survey, 1954–1971* Tokyo: Nihon Tokei Kyokai, 1955–72

Japan Ministry of Agriculture and Forestry *Statistical Yearbook of the Ministry of Agriculture and Forestry 28–48* Tokyo Norin Tokei Kyokai, 1952–73

———. *Milk and Dairy Product Statistics, 1970* Tokyo. Norin Tokei Kyokai, 1971.

Japan Ministry of Finance *Japan Exports and Imports, 1967–1970.* Tokyo Japan Tariff Association, 1968–71

Ophir, Tsvi "The Interaction of Tariffs and Quotas" *A E R* 59 (December 1969) 1002-5.

A Purchasing-Power-Parity Application to Demand for Money in Afghanistan

Maxwell J Fry

University of Hawaii

This paper presents an estimate of the demand for money in Afghanistan. The basic problem is the lack of reliable price data.[1] However, Da Afghanistan Bank (the central bank) has recorded the free-market foreign exchange rates daily since 1951 [2] They are used to overcome this difficulty through an application of the purchasing-power-parity theory to construct a domestic price index

The purchasing-power-parity theory states that

$$p = rw, \tag{1}$$

where p is the domestic price index, r the exchange rate, and w the price index in the country whose currency is used for the exchange rate. In this case, r is the Afghani/dollar rate and w the U S wholesale price index The Indian and Pakistani rupee rates might have been chosen as alternatives to the Afghani/dollar rate The advantage would have been a closer correspondence of consumption patterns The disadvantage, however, is that there has been extensive price control in both of these countries as well as abundant foreign trade restrictions. Initial experimentation produced poor results using these rates and the corresponding price indices Other convertible currency rates could have been calculated from the Afghani/dollar rate Given Afghanistan's trading pattern and the United States' predominance in the international economy, no advantage could be expected from so doing

---

I would like to thank Jacob A Frenkel, Moheb Ghali, Harry G Johnson, Deena Khatkhate, H Laurence Miller, Stig von Post, and John H Power for their comments on earlier drafts of this paper

[1] The official domestic price index covers only the past decade and is extremely inaccurate It was the first task of the Afghan Central Statistics Office, established in 1972, to prepare a new price index Revisions of the official index for three previous years produced by the Central Statistics Office in one of its first publications (1972) indicate how misleading the series for the 1960s is

[2] Monthly averages of the daily dollar, Indian rupee, and Pakistani rupee exchange rates were provided by the Research Department of Da Afghanistan Bank (Fry 1974, tables 7 20, 7 21, 7 22)

[*Journal of Political Economy*, 1976, vol 84, no. 5]
© 1976 by The University of Chicago. All rights reserved

Balassa (1964) showed that the purchasing-power relationship does not imply that the cost of living will be the same in Afghanistan and the United States. A wide disparity in levels of productivity between these two countries results in much lower prices of nontraded goods and services in Afghanistan than in the United States. For example, in 1971 full-time domestic servants employed by British Embassy personnel were paid about $22 a month.[3] However, as long as rates of change in labor productivity in Afghanistan and the United States have not differed greatly and the proportion of traded to nontraded goods in both countries has remained unchanged over the period under analysis, the relative version of the purchasing-power-parity theory may hold. This version states that the rate of change in domestic prices will be the same as the rate of change in the purchasing-power-parity index.

Except during World War II, the foreign exchange rate has always been determined by the free interplay of market forces. Apart from a few unimportant items prohibited mainly for religious reasons, there have been no restrictions on foreign trade. No major disruptions to the economy have occurred since World War II. The foreign exchange markets (known as money bazaars) of Kabul and Kandahar operate competitively on fine margins. The range of activities in the money bazaars—an accepted institution in the life of businessmen and traders in Afghanistan—is impressive. Hence, the hope that the purchasing-power-parity theory has been valid in Afghanistan's case over the postwar period may not be too unrealistic. Initial econometric tests of various exchange-rate and purchasing-power-parity equations were all consistent with this theory.[4]

The demand for money function estimated for the period 1952–71[5] takes the form

$$\ln (M/p) = a_0 + a_1 \ln (y) + a_2 i^*, \tag{2}$$

where M is aggregate nominal money balances broadly defined to include time deposits, p is the domestic price level as given by the purchasing-power-parity relationship in equation (1), y is real monetized income, and i^* is the expected rate of inflation, taken to be the opportunity cost of holding money in Afghanistan.[6] The latter is derived as follows:

$$i_t^* = \sum_{j=0}^{n} b_j \frac{\Delta(r_{t-j} w_{t-j})}{r_{t-j-1} w_{t-j-1}} \tag{3}$$

[3] Using Balassa's equation $Y = 49.34 + 0.025X$ (Balassa 1964, p. 590), a per capita income of $100 implies that the absolute purchasing-power-parity price index overestimates the price level in Afghanistan by a factor of almost two.

[4] A longer version of this paper in which these tests are fully reported is available on request from the author (see also Fry 1976).

[5] The starting date was determined by data availability, and the concluding date by the devaluation of the dollar in 1971.

[6] There are virtually no financial assets other than money. Institutional interest rates remained low and unchanged over the period.

The coefficients, b_j, are estimated using the Almon polynomial lag technique. Government domestic revenue deflated by the purchasing-power-parity price index, g, is used as a proxy for real monetized income. It is a reasonable proxy, since there have been no major tax changes over this period, tax effort has been consistently low (Fry 1974, p. 182). The basic data used in the regression analysis are given in table 1.

Equation (4) was obtained from two-stage least-squares estimation to eliminate simultaneous-equation bias[7] and was corrected for serial correlation using the Cochrane-Orcutt technique.[8]

$$\begin{array}{lcccc} \ln (M/rw) = & 1.822 + & 0.828 \ln (g) & - & 1.059\, \hat{\imath}^{*}, \\ \text{SE} & (0.446) & (0.059) & & (0.158) \end{array} \qquad (4)$$

$$R^2 = .937,\ \text{SEE} = 0.024,\ \text{D-W} = 2.397,\ \rho = -0.374$$

The rate of inflation, i, was estimated in the first stage and the estimated values, $\hat{\imath}$, used in equation (4). The lagged coefficients for $\hat{\imath}^*$ obtained from an unconstrained polynomial of third degree are $\hat{\imath}_t = 0.080$, $\hat{\imath}_{t-1} = 0.087$, $\hat{\imath}_{t-2} = 0.127$, $\hat{\imath}_{t-3} = 0.174$, $\hat{\imath}_{t-4} = 0.206$, $\hat{\imath}_{t-5} = 0.198$, and $\hat{\imath}_{t-6} = 0.128$. The length of the lag was chosen on a nonnegativity criterion.

The variation in $\ln [M/(rw)]$ is well explained by the two independent variables. Coefficients have the correct signs and economically meaningful values.[9] The lag structure is of the anticipated inverted U shape.

The coefficient of real domestic revenue suggests that the elasticity of

---

[7] The instrumental variables used were first differences of the logarithms of government expenditure, exports, foreign aid receipts, all at constant prices, and the nominal money stock. All variables, including the dependent variable, were also introduced with a 1-year time lag. Because of the small sample size, there is some question as to the appropriateness of the two-stage least-squares procedure. The equation estimated by ordinary least squares is

$$\begin{array}{lcccc} \ln (M/rw) = & 1.971 + & 0.810 \ln (g) & - & 1.047\, i^{*}, \\ \text{SE} & (0.241) & (0.032) & & (0.084) \end{array}$$

$$R^2 = .976, \quad \text{SEE} = 0.015, \quad \text{D-W} = 2.807, \quad \rho = -0.489.$$

It is almost identical to eq. (4) in the text.

[8] Without correcting for simultaneous-equation bias and serial correlation, lag structures tended to exhibit a V shape. This same problem was encountered by Dickson and Starleaf (1972, p. 1037) and Goldfeld (1973, table 8) and was eliminated by the former through correction for first- and second-order serial correlation. This procedure invalidates the usual interpretation of standard errors and the coefficient of determination. They are reported here simply as rough guides with which to assess correlation.

[9] The coefficient of i^* and the mean inflation elasticity (0.060) are comparable to those found in a 47-country analysis by Perlman (1970, pp. 315, 336–37) and in estimates for the neighboring countries of Iran and Pakistan (Fry 1973, pp. 377–79, Abe et al. 1975, p. 255). Studies of hyperinflation using the semilog equation generally produce larger coefficients and much higher mean elasticities.

TABLE 1

Basic Data Used in Regression Analysis

Year	Money Stock ($M2$)*	Government Domestic Revenue*	Afghani/ Dollar Exchange Rate	U.S. Wholesale Price Index (1961 = 100)	Purchasing-Power-Parity Price Index (1961 = 100)	Rate of Inflation, 100π (%)	Estimated Rate of Inflation, $100\hat{\pi}$ (%)
1952	1,274	614	37.0	93.7	79.7	2.71	
1953	1,339	775	37.1	92.4	78.8	−1.13	
1954	1,593	964	42.1	92.6	89.6	13.71	10.97
1955	2,097	1,283	45.2	92.9	96.5	7.70	9.21
1956	2,473	1,149	53.8	95.9	118.6	22.90	22.72
1957	2,481	1,348	54.1	98.7	122.8	3.54	3.38
1958	2,417	1,533	54.7	100.1	125.9	2.52	2.83
1959	2,587	1,601	46.6	100.3	107.4	−14.69	−11.95
1960	2,728	2,099	40.8	100.4	94.2	−12.29	−13.73
1961	2,981	2,110	43.5	100.0	100.0	6.16	3.38
1962	3,565	2,123	52.8	100.3	121.7	21.70	16.89
1963	4,405	2,677	51.3	100.0	117.9	−3.12	−2.67
1964	5,393	3,061	63.6	100.2	146.5	24.26	28.32
1965	6,160	3,976	75.3	102.2	176.9	20.75	18.73
1966	6,355	4,285	76.4	105.6	185.5	4.86	1.83
1967	6,289	4,211	76.4	105.8	185.8	0.16	1.80
1968	6,538	4,465	74.8	108.4	186.4	0.32	0.83
1969	7,341	5,085	75.4	112.7	195.3	4.77	6.66
1970	8,181	5,702	84.8	116.8	227.7	16.59	17.72
1971	8,759	5,823	84.6	120.6	234.5	2.99	5.91

Source.—Fry (1974, tables 2.6, 2.9, 2.13)

Note.—The money stock series are averages of beginning- and end-of-year figures.

* In millions of Afghanis (current prices)

demand for money has been less than unity over this period.[10] This might seem surprising for a less-developed country However, independent evidence can be introduced to corroborate it Income elasticity determines movements in velocity as income changes If elasticity is less than unity, velocity increases with income For less-developed countries, a significant trend in velocity is usually accompanied by a trend in the same direction in the currency/money ratio (Khazzoom 1966, 1967) *Ceteris paribus*, both can be expected to fall during the course of economic development (Cagan 1958). However, the currency/money ratio in Afghanistan rose from a low of 0.64 in 1955 to 0 81 in 1962, only declining to 0 76 by 1971 (Fry 1974, table 2.12) The higher velocity implied by higher real domestic revenue (and hence income) at the end of the period is thus matched by a higher currency/money ratio[11]

Afghanistan differs from two groups of countries for which demand for money estimates abound It is not a rich, industrial country, nor has it suffered from hyperinflation Afghanistan is, in fact, one of the poorest countries in the world The annual rate of inflation between 1952 and 1971, although highly volatile, averaged only 5 7 per cent The results reported above strongly suggest, however, that Afghanistan does share a common characteristic with these other two groups the estimated demand for money, determined in predictable fashion by income and cost, is stable

References

Abe, S , Fry, M J , Min, B K , Vongvipanond, P , and Yu, T -P "The Demand for Money in Pakistan Some Alternative Estimates " *Pakistan Development Rev.* 14 (Summer 1975) 249-57

Balassa, B "The Purchasing-Power Parity Doctrine A Reappraisal " *J.P.E.* 72, no 6 (December 1964) 584-96

Cagan, P "The Demand for Currency Relative to the Total Money Supply." *J P E* 66, no 4 (August 1958) 303-28

---

[10] It is assumed that the proportion of monetized income in total income has been constant over this period The realism of this assumption is discussed at some length in connection with monetary estimates of gross national product presented in Fry (1974, pp 29-42). From the scanty data available, it does seem that this proportion has not changed to any significant extent over the period 1952-71

[11] The explanation for both lies in the institutional monetary history of the period A banking boom took place during the early 1950s Within a decade, one state-owned commercial (Pashtany Tejaraty Bank) and three state development banks had opened their doors alongside Bank Millie (the only private commercial bank) and Da Afghanistan Bank. By 1957 the fortunes of this infant financial sector were already on the decline. Political interference, weak management, and legal impediments to securing loans, in addition to inflationary conditions over the years 1954-58, resulted in a high rate of attrition. The development banks had all virtually ceased operating by the end of the 1950s Pashtany Tejaraty Bank maintained only a tenuous existence which any significant net deposit withdrawals would have seriously jeopardized A year after its establishment, the observation was made that "*it is hard to see how it can earn enough to pay dividends*" (Wilbur 1956, p 185).

Central Statistics Office. "Revised National Price Indexes for Afghanistan for the Years 1348–1350." Mimeographed. Kabul: Central Statis Office, 1972.

Dickson, H. D., and Starleaf, D. R. "Polynomial Distributed Lag Structures in the Demand Function for Money." *J. Finance* 27 (December 1972): 1035–43.

Fry, M. J. "Manipulating Demand for Money." In *Essays in Modern Economics*, edited by M. Parkin with A. R. Nobay. London: Longman, 1973.

———. *The Afghan Economy: Money, Finance and the Critical Constraints to Economic Development*. Leiden: Brill, 1974.

———. "A Monetary Approach to Afghanistan's Flexible Exchange Rate." *J Money, Credit, and Banking* 8 (May 1976): 219–25.

Goldfeld, S. M. "The Demand for Money Revisited." *Brookings Papers Econ. Activity* 3 (1973): 577–638

Khazzoom, J. D *The Currency Ratio in Developing Countries*. New York· Praeger, 1966.

——— "Covariations in the Currency Ratio and the Velocity of Money in Underdeveloped Countries" *J Development Studies* 3 (April 1967) 293–306

Perlman, M. "International Differences in Liquid Assets Portfolios" In *Varieties of Monetary Experience*, edited by D. Meiselman. Chicago Univ. Chicago Press, 1970.

Wilber, D N., ed. *Afghanistan* New Haven, Conn Human Relations Area Files, 1956.

Book Reviews

Gold, Money and the Law. Edited by HENRY G. MANNE and ROGER LEROY MILLER. Chicago: Aldine Publishing Co., 1975. Pp. 220 $10.75 (cloth).

Franklin D Roosevelt's gold policies continue to exercise our curiosity even after a lapse of more than 40 years The year 1933 was a year of considerable monetary experimentation and uncertainty. The president suspended gold payments, nationalized the gold stock, abrogated the gold clauses, and bought gold in both domestic and foreign markets. Finally, in January 1934 he devalued the dollar. The occasion for the revival of interest in this episode was the recent lifting of the ban on private gold holdings by U S. citizens and the renewal of interest in the problem of indexation as a defense against continued inflation.

The Center for Studies in Law and Economics at the University of Miami School of Law (Florida) convened its inaugural conference to discuss both the legal and economic implications of the famous *Gold Clause Cases* in which the Supreme Court upheld congressional abrogation of all such clauses in private contracts The topic was selected presumably to demonstrate how law and economics "can be integrated in grappling with major legal and economic issues " But the real issues raised at this conference go far deeper than these matters of current and historical interest Gold is merely the pretext for a cursory exploration of the questions, What are the fundamental limits of the government's monetary authority? What sort of monetary constitution might be predicted to emerge from a genuine social contract? And did our monetary institutions before 1933 bear any relationship to this ideal structure? These are indeed important questions and have not received the attention they deserve This book is an invitation to remedy that deficiency

The principal paper of the conference was presented by James Buchanan and T Nicolaus Tideman. Formal replies were made by two economists, Milton Friedman and Harry Johnson, and two lawyers, Ralph K. Winter, Jr., and Gerald T Dunn Following the formal papers there was an extended discussion by the more than 50 conference participants which is fully reported in the book. A 70-page initial paper which evoked twice as many pages of comment should be testimony either to the contentiousness of the participants or to the intellectual excitement created by the authors. My comments will be restricted to the economic rather than the legal issues raised, not because they are less interesting but because I feel less qualified to deal effectively with constitutional questions.

Buchanan and Tideman define a contractarian monetary framework in terms of two separate functional objectives (1) definition of the monetary unit and (2) predictability in the value of money Granted that predictability in the value of money is a desirable attribute of whatever monetary arrangements prevail, which is a far cry from enshrining that attribute alone in a monetary constitution. Conceivably, there are occasions when predictability of value might have to be sacrificed to attain other social goods which are accorded a higher priority, most notably in wartime.

Buchanan and Tideman apply their hypothetical contractarian framework in an attempt to answer the question, Was there any economic justification for the suspension of convertibility and abrogation of the gold clauses in private and public contracts? They find a rationale in modern welfare economics by appealing to the public-goods aspect of money and the existence of externalities. For example, they maintain that there is a valid contractarian argument for FDR's suspension of convertibility if the risks were great that the public would reject national fiat money and search for ways to substitute gold in exchange transactions, thereby making the exchange process, in their judgment, less efficient. Friedman correctly pointed out that on the purely factual level the possibility of this having happened was indeed remote because there was little or no distrust of the national currency (as opposed to confidence in bank-created money) and the costs associated with a dual currency system were probably minimal

Moreover, Buchanan and Tideman find justification for the suspension of gold clauses in private contracts during periods of unforeseen emergencies when the relative positions of debtors and creditors are shifted unexpectedly The enforcement of the gold clauses in contracts would have required at the time an increase in payments in real terms by almost 70 percent, unquestionably seriously damaging debtor interests.

Was there a viable alternative to devaluation and expropriation of gold holdings as deduced by Buchanan and Tideman from the contractarian framework? Friedman and Johnson think that there was. All that the Federal Reserve and FDR had to do was to increase high-powered money in the form of Federal Reserve notes and deposits The increase in high-powered money would have prevented the reduction in the total quantity of money and ended bank failures. According to Friedman and Johnson, the Federal Reserve had the knowledge and previous experience to have acted in the manner prescribed. But there was monumental ineptness and mismanagement on the part of Reserve officials between 1929 and 1933 I find little or no basis for the Friedman-Johnson indictment of Federal Reserve monetary policies in my own studies of the same period. They have overstated the importance of short-term countercyclical objectives in the formation of monetary policy in the twenties and thirties and exaggerated the extent to which the System understood the use of open market policy to counteract financial or any other kind of economic instability

Nevertheless, I agree with Friedman that there is little evidence to support Buchanan and Tideman's conjecture that continued adherence to the gold standard would have prevented or delayed necessary monetary expansion Such a consideration never entered Roosevelt's calculations about suspending gold payments Moreover, the postdevaluation gold inflow was only partly due to the alteration in the exchange rate A considerable part was attributable to political uncertainty in Europe and the rise of Hitler.

Although Buchanan and Tideman have provided us with an interesting theoretical rationale for explaining the events of 1933, I fail to see how it contributes to our understanding of the economic history of the period Conjectural history is no substitute for a patient and detailed investigation of FDR's gold policies. Admittedly, that was not the purpose of Buchanan and Tideman's paper. Less heat might have been generated by their critics had the authors been less sanguine about the historical relevance of their hypothetical analysis. Furthermore, I am not convinced that the Supreme Court decision to uphold the abrogation of the gold clauses in 1933 tells us anything about how the Court might view the validity of index contracts in general today.

ELMUS R. WICKER

Indiana University

Essays in Monetary Theory and Related Subjects. By JØRGEN PEDERSEN. Copenhagen. Samfundsvidenskabeligt Forlag, 1975. Pp. 233.

Jørgen Pedersen was an original and forceful thinker. Like Carl Föhl and Albert Hahn, he developed independently many of the ideas which became attached to the name of Keynes, but he also shared that skepticism about short-run monetary policy we now associate with the name of Friedman These essays show him as an outstanding macroeconomist with a sure sense for the substantive economic issues and the courage to follow his own arguments to their logical conclusion (which was wage control). Since his work is largely unknown to American economists, it may be appropriate to use this review for a summary of his ideas as they emerge from these essays

Business fluctuations are seen to arise from shifts between relatively rigid prices (particularly wages) and flexible prices (particularly international prices) The main task of economic stabilization is to prevent such shifts To force wages to adjust to international prices can only result, Pedersen argued in 1933, in disastrous fluctuations in employment He advocated the adjustment of international prices to domestic wages through variable exchange rates combined with monetary and fiscal policies designed to keep the level of money incomes constant.

In 1937 Pedersen developed the principles of "functional finance" that flow from this view of stabilization There is no particular virtue in balanced budgets, and whether expenditures are financed by taxes, debt, or money depends on business conditions These conditions are already perceived in terms of the dichotomy between unemployment with fixed prices and full employment with flexible prices that should plague Keynesian theorizing to the present day. Unemployment required money financing while tax financing was appropriate for full employment There should be few occasions for debt financing, because the "crowding out" of private investment through higher interest rates makes it undesirable in a depression, and under full employment tax financing is less inflationary

Pedersen's excursions into monetary theory mainly show that he, like many Keynesians, failed to understand the significance of the money stock (as opposed to the flow of money expenditures) for the long-run trend of prices Inflation thus became for him a problem of money wages, with the gold standard supplanted by a wage standard As an early and ardent advocate of an "incomes policy" through wage controls, he was caught in a familiar dilemma Since relative wages had to be flexible for allocative purposes and absolute wages could, he thought, never fall, some degree of inflation seemed to be inevitable, its degree depending on "public education."

But Pedersen was too independent to be simply a Keynesian The two later essays on the stabilization of the mark and the causes of the Great Depression (the best and freshest reading in the collection) show that the main basis for his distrust of monetary policy was not doubt of its effectiveness but the conviction that the disastrous consequences of ill-timed and excessive monetary contraction would far outweigh the potential benefits With the belief that serious economic crises are caused not by autonomous fluctuations in investment demand but by errors in monetary policy, Pedersen is closer to Friedman than to the later Keynesians.

Pedersen was also a strong believer in the allocative function of markets He criticized credit rationing and variable reserve requirements as inefficient instruments of monetary policy, producing a misallocation of credit and sharp liquidity crises. Commercial banks, he argued, should be able to borrow unlimited amounts at the market rate, but this rate should be varied through open market operations. However, monetary policy should be geared to long-run objectives, while the short-run "crisis management" ought to be left to fiscal policy, exchange rate adjustments, and international cooperation of central banks (whose power Pedersen overestimated in the early thirties in the same way central bankers overestimated it in the early sixties).

Most of what Pedersen had to say has been said better or more effectively, and "incomes policy" has been a consistent failure. It is impressive to see, however, how an independent mind, having inherited largely the same analytical tools as Keynes and being confronted with the same economic experience, arrived at essentially the same policy conclusions. The American economist tends to view economic science as epitomized in the works of a few great men writing in English. Pedersen's collected essays should be enough to correct this error.

JÜRG NIEHANS

Johns Hopkins University

Selected Economic Essays and Addresses By ARNOLD PLANT
London Routledge & Kegan Paul, 1974 Pp. xi + 248.

Many other economists, I suspect, suffer from the same lack of knowledge of Sir Arnold Plant s writings as I did prior to finding *Selected Economic Essays and Addresses* on my desk If this is true, it is clear that we have neglected an important figure in relatively recent thought.

Sir Arnold, born in 1898, was a mechanical engineer during and after World War I For reasons unclear from brief biographical notes, he became a student at LSE in 1920, graduated with bachelor of commerce and bachelor of science (economics) degrees in 1923, and took an appointment at the University of Cape Town, South Africa, in 1924 He accepted a chair at LSE in 1925, from which he retired in 1965. His knighthood was conferred in 1947

The volume considered here is divided into three parts, "African Studies," "Property and Ownership," and "Economic Processes at Work." The eleven essays and addresses were prepared between 1927 and 1970 and represent but a small part of his lifework There is an even, persistent theme—one which might be described as Hayekian or, after one of his better-known students, Coasian. There is a persistent method of approach, too, of discussing a subject in deliberate detail, noting alternative views, and then coming to rather definitive conclusions.

Consider the first chapter, "The Economics of the Native Question " After polite yet biting critique of legislation affecting the black natives, with an aim at affecting proposed amendments, Sir Arnold, with good use of economic analysis, concluded. "If the agrarian revolution contemplated . . . is not to create an army of discontented natives, to dislocate farming operations, and cause serious repercussions in the towns, its provisions must be amended so as to convert squatters by an inevitable, *gradual* process into peasant proprietors or leaseholders on the one hand, and into contented labourers on the other" (p. 16) To cap this, on the issue of education and voting rights, Sir Arnold pointed out: "It is relevant to point out that it is the Europeans who must bear the responsibility for withholding civilisation from the natives, but if [it is suggested] that fitness for the franchise, and capacity to maintain and further Western civilisation are inherited characters, it will find little support among modern biologists" (pp. 16–17). How popular these views must have been to whites in South Africa—most of whom could read but probably didn't—as well as to the civil servants in London—most of whom could read and some of whom probably did!

A review can hardly be as long as the work reviewed Subsequent chapters are "Economic Issues in Immigration" (1970), "The Economic Theory concerning Patents for Inventions" (1934), "The Economic Aspects of Copyright in Books" (1934), "The New Commerce in Ideas and Intellectual Property" (1953), "The Substance and the Shadow—Reflections on Prosperity" (1956), "Engineering in an Expanding Economy" (1959), "Homage to Hayek" (1970), "Centralise or Decentralise" (1937), "The Distribution of Proprietary Articles" (1937), and "The Economic Approach to the Peaceful Use of Nuclear Power" (1958). From these, a few more quotations.

> As I see it, therefore, the prolongation of copyright payments and of the consequent higher prices . . are not economically necessary. [Pp. 93–94]
>
> People, then, differ widely in the values they set on the various constituents of well-being Lord Snowden . . once remarked that it was a mistake to assume that the people of this country would be satisfied with full employment, guaranteed food and shelter, and an adequate medical service . [P. 124]
>
> I have allowed myself to be carried away. . . [P. 125]
>
> It seems clear, first, that if the poorer countries wish above all else to raise real incomes, the best general policy for most of them will be . to expand their production of staple commodities for export. . . . It is, however, true that the world prices of their exports . . fluctuate more widely than the world prices of their imports . . The richer countries should no longer lay themselves open to the old charge that, whenever the United States, or Western Europe or the United Kingdom sneezes, they see to it that the rest of the world catches pneumonia. . . [I]t is not surprising that [poorer countries] should seek to obtain greater internal stability by diversifying . . beginning to manufacture some of the commodities which they have hitherto imported. [Pp. 156–57]

If I have succeeded in encouraging just a few to leaf through Sir Arnold's *Essays and Addresses*, my purpose has been accomplished. It is a fascinating, pleasant, and rewarding experience

ALMARIN PHILLIPS

University of Pennsylvania

ENVIRONMENTAL IMPACT ASSESSMENT/GROWTH MANAGEMENT/ AND THE COMPREHENSIVE PLAN
Joseph Lee Rodgers, Jr., University of Oklahoma
Available $15 00

TAX SHELTER SALE-LEASEBACK FINANCING
The Economic Realities
Stephen E Roulac, University of California, Berkeley
Available $20 00

NONMETROPOLITAN URBAN HOUSING An Economic Analysis of Problems and Policies
Michael A Stegman, The University of North Carolina at Chapel Hill and Howard J Sumka, University of Kansas
Available $16 50

ELECTRIC ENERGY AVAILABILITY AND REGIONAL GROWTH
George S. Tolley and Charles W Upton, University of Chicago and V Stevens Hastings Associates for Economic Analysis
In preparation ca $17 50

FEAR IN THE COUNTRYSIDE The Control of Agricultural Resources in the Poor Countries by Non-Peasant Elites
E G Vallianatos, Center for Population Studies, Harvard University
Available $14 00

PRISON REFORM AND STATE ELITES
Richard A Berk, Northwestern University and Peter H Rossi, University of Massachusetts
In preparation ca $15 00

CONSTRAINTS ON THE REGULATORY PROCESS: A Case Study of Regulation of Cable Television
Richard Olin Berner
(Published for the Program on Information Technologies and Public Policy, Harvard University)
In preparation ca $13.50

RESISTANT PR'CES: A Study of Competitive Strains in the Antibiotic Markets
Paul A Brooke
Foreword by Robert Heilbroner
(Published for the Council on Economic Opportunities)
Available $12 50

THE FUTURE OF WORLD OIL
Paul L Eckbo, Energy Laboratory, Massachusetts Institute of Technology
Available $15 00

POLICING A FREE SOCIETY
Herman Goldstein, University of Wisconsin Law School
Available $15 00

PRIMARY CARE IN A SPECIALIZED WORLD The Failure of Health Manpower Politics
Philip R Lee and Lauren LeRoy, University of California, San Francisco, Janice Stalcup, City University of New York, and John Beck, University of California, San Francisco
Foreword by The Honorable Paul G Rogers
Available $15 00

NEW COMMUNITIES RESEARCH SERIES

Center for Urban and Regional Studies of The University of North Carolina at Chapel Hill, Shirley F Weiss and Raymond J Burby III, Series Editors

This series summarizes the results of a nationwide study and evaluation of new community development in the United States The following titles are available Fall, 1976, additional titles will be published early in 1977

ECONOMIC INTEGRATION IN NEW COMMUNITIES
An Evaluation of Factors Affecting Policies and Implementation
Helene V Smookler ca $15 00

RECREATION AND LEISURE IN NEW COMMUNITIES
Raymond J Burby, III ca $17 50

RESIDENTIAL MOBILITY IN NEW COMMUNITIES
An Analysis of Recent In-movers and Prospective Out-movers
Edward J Kaiser ca $15 00

Available on 30-day approval from Ballinger

Ballinger
PUBLISHING COMPANY
17 Dunster Street • Cambridge, Mass. 02138

A New and Superior Process for Making Social Choices

T. Nicolaus Tideman and Gordon Tullock

Virginia Polytechnic Institute and State University

This paper describes and elaborates a process first discovered by Edward H. Clarke that motivates individuals to reveal their true preferences for public goods. The essence of the process is that each individual is offered a chance to change the outcome that would occur without his vote by paying a special charge equal to the net cost to others that results from including his vote in the decision. Because the special charge on any one person is not paid to any other person, a very small budget surplus results. Applications to both discrete and continuous decisions are illustrated.

This paper describes a new process for making social choices, one that is superior to other processes that have been suggested. The method is immune to strategic maneuvering on the part of individual voters. It avoids the conditions of the Arrow theorem by using more information than the rank orders of preferences, and selects a unique point on or "almost on" the Pareto-optimal frontier, one that maximizes or "almost maximizes" the consumer surplus of society. Subject to any given distribution of wealth, the process may be used to approximate the Lindahl equilibrium for all public goods.[1]

These are strong claims, and it is therefore only sensible to begin this paper by pointing out that the process will not cure cancer, stop the tides, or, indeed, deal successfully with many other problems. As far as we know, all existing social-choice processes are subject to exploitation by suitably designed coalitions. This process is no exception. In addition, as in all democratic voting processes, voters are undermotivated to

---

[1] The method is also applicable to decisions about income and wealth redistribution, instead of leaving that issue aside in the conventional manner. It can be used to ensure the competitiveness of markets, and it provides a welfare criterion superior, in our opinion, to Pareto optimality. All of these matters must be deferred for later publication.

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved.

invest time and effort in a comparative evaluation of alternatives. The motives they are given for making sensible decisions are somewhat stronger than those for persons engaging in voting under majority rule, but voters will be asked to do more in the way of expressing their preferences than simply saying yes or no Therefore, it is not clear whether the lack of an incentive to vote is more or less of a handicap for this process than it is for ordinary voting processes

The process may be described most generally as a demand-revealing process It relies on what might be called an incomplete compensation mechanism that appears to have been first described by Vickrey (1961) in the context of optimum counterspeculation policy for a socialist economy. The essence of the mechanism is that each person is paid for the benefits (or pays the costs) of his actions, but no one is charged (or credited) as required for budget balance Vickrey showed that it would be possible to motivate individuals to reveal their true supply and demand schedules for a private good by paying each person the net increase in the sum of the producer and consumer surpluses of other persons in the market that resulted from the supply or demand schedule that the one person reported. Vickrey noted that there would be a problem of financing such a system, since it would generate a deficit. He did not discuss the potential application of such a system to public goods

Two persons who were unaware of each other's work or Vickrey's discovered the applicability of a similar compensation mechanism to the problem of motivating individuals to reveal their true demands for public goods The first to publish was Edward Clarke (1971, 1972), whose papers until now have made very little impact on the economics profession The lack of impact can be attributed partly to the nature of the idea that Clarke put forward, which is counterintuitive to almost any welfare economist, and partly to Clarke's difficult writing style. The second person was Theodore Groves. In one paper (Groves 1973) he offered a mathematically rigorous treatment of a procedure like Vickrey's for allocating scarce private goods within an organization. More recently Groves and Loeb (1975) published a procedure isomorphic to Clarke's for selecting the optimal quantities of public goods. Our objectives in this paper are to provide a clear explanation of the demand-revealing process as it applies to public goods and to extend the understanding of the process on several fronts.

While, as Bowen (1943) showed, majority rule is efficient if the intensity of voters' preferences is distributed symmetrically, the demand-revealing process does not require for its efficiency any restriction on the distribution of intensities of voters' preferences. Unlike the voting processes proposed by Thompson (1965), Dreze and de la Vallee Poussin (1971) and Tideman (1972), the demand-revealing process requires no special beliefs on the part of voters. Basically, it provides an environment in which each voter is motivated to reveal his preferences correctly. This is accomplished by the use of a special—indeed, bizarre—tax mechanism which rewards truthful presentation of preferences and penalizes concealment or falsification.

In order to explain the process, we start not with the problem with which Clarke's first paper dealt, the choices of the optimal amount of a single-dimensional public good, but with the simpler case of a choice among discrete options, which Clarke's second paper discussed cryptically. For simplicity, we shall start with two alternatives, which may be conceived of as two policies or two candidates We shall then show how the process can be extended to more than two options Having introduced the subject with these simple examples, we shall then turn to the choice of the optimal amount of a public good.

Choice between Two Options

Suppose that a collective choice must be made between two options, designated A and B The rule we describe involves asking each individual to state which option he prefers and the amount of money he is willing to pay to secure his preferred option instead of the other We shall show shortly why he would have an incentive to respond truthfully. In table 1, we show the "votes" of each of three voters for the two options Option A is worth a total of $70 to the persons who prefer it and is chosen by the rule because B is worth less to its proponent.

We now turn to why the voter is motivated to correctly state his preferences There is a "Clarke tax" to be levied, and, as we said before, it is a bizarre tax. We inquire with respect to each voter what the outcome would have been if he had not voted For example, if voter 1 had not voted, then the outcome would have been that alternative A would have received $40 total and alternative B $60; hence, alternative B would have won. We charge voter 1 $20, the amount necessary to bring the "votes" for A up to equality with the "votes" for B. By the same line of reasoning, voter 3 pays a tax of $30 Voter 2 pays no tax because his vote did not change the outcome Note that, had voter 1 understated his preference for A by an amount less than $10, he would have paid exactly the same tax as he did. If he had understated his preference for A by more than $10, B would have been selected. And voter 1 would prefer having A at the price of $20 to having B. Similarly, if a voter overstates his preferences, either the overstatement makes no difference in what is selected or what he pays or else (e.g., if voter 2 said B was worth $100 to him) he changes the result by his action and pays more for his choice than it is worth to him.

To describe the decision rule generally, define S_A as the sum over all voters who state a preference for A over B of the amounts they offer to

TABLE 1

AGGREGATING PREFERENCES FOR TWO OPTIONS (IN DOLLARS)

VOTER	DIFFERENTIAL VALUES OF OPTIONS A	B
1	30	0
2	0	60
3	40	0
Total	70	60

pay to have A instead of B. Define S_B similarly. The collective-choice rule will be to choose A if $S_A > S_B$, choose B is $S_B > S_A$, and flip a coin to decide if $S_A = S_B$ The incentive to respond truthfully is generated by a "Clarke tax," a rule that a voter must pay a portion of his offer if and only if his vote changes the outcome Any voter who changes the outcome must pay $|S_A - S_B|$, calculated without his vote In the case of a tie that is decided by a coin toss, every voter on the side that wins the toss is regarded as having changed the outcome. If the result without a person's vote is a tie, he pays nothing.

In effect, this rule gives each voter the choice of (1) leaving the outcome where it would be without his vote or (2) changing it at a price of the reported net loss to other voters. If the value to a voter of his preferred outcome is less than the net value of the alternative to others, then he prefers (1), which occurs if he responds truthfully If his value is greater than the aggregate net value to others, then he prefers (2), which again is what occurs upon a truthful response. If his value exactly balances the net value reported by others, then he is indifferent between the two possibilities, and we flip a coin if he responds truthfully. A nontruthful response cannot benefit the respondent, and it carries a risk of making him worse off than he would have been with the truth. If he understates his value, he may pass up an opportunity to obtain the result he desires at an attractive price If he overstates his value, he may wind up paying more than it is worth to him to have his choice.

To characterize the rule in terms of property rights, one might call it "entitlement to the consequence of one's abstention," since the result that occurs if a voter abstains costs him nothing; while if his vote changes the collective choice, he must pay. This has a certain family resemblance to majority rule, where the voter's only entitlement is to what a majority of persons other than himself want, except that, if all others tie, he can decide the issue.

Any money collected from voters in this system must be wasted or given to nonvoters to keep the incentives correct. If voters received the money collected, the possibility of increasing their shares would distort their incentives. However, if the revenue were simply divided equally among all voters, the effective distortion would be minimal if there were more than 100 or so voters; with a large number of voters, it is most likely that no one vote will change the outcome, so that in most cases no taxes for voting will be collected. We will discuss the significance of the lack of budget balance in more detail in the context of decisions about continuous variables

A serious problem in all voting systems is the weakness of the incentive to vote. The demand-revealing process is no exception. The only wholly instrumental reason for a person to vote is the possibility that his vote will be decisive Failure to vote carries a risk of passing up a chance to alter the outcome at a favorable price, but the probability of being decisive is usually small enough so that people might still reasonably conclude that voting is not worth the effort Even if people do decide to vote, they are normally not motivated to give any serious study to their vote in collective decision processes, because the probable gain from acquiring further information or simply reflecting on the information already at hand is usually less than the cost Thus, ill-informed voting is to be expected. The demand-revealing process is no exception to this general rule put forward by Downs (1957).

It may seem that a person who sustains a large loss when his preference is not followed deserves compensation, but this cannot be given without motivating an excessive statement of differential value If a voter expected to lose, an offer to compensate his loss would motivate a statement from him of a larger loss. In regard to the uncompensated losses that are produced, the demand-revealing process is similar to majority rule. In the latter, every voter must live with the choice of the majority. His only opportunity to be decisive on the issue is if there is an equal number of other voters on each side. Similarly, in the demand-revealing process, every voter must live with a finding that the aggregate value placed by others on the alternatives is opposed to his own interest, provided that he is not willing to pay enough to give his preference the higher aggregate value.

It might be objected that the demand-revealing process would permit confiscatory action. If there is a proposal to tear down one person's house and make the site a park, and if others report a greater gain from the park than the occupant's loss, then the latter loses his property. The demand-revealing process would indeed have this confiscatory characteristic if there were no constitutional limits on the proposals that could be considered. In this respect, the system is again like majority rule, which has a similar confiscatory potential. It is reasonable to expect that people making collective choices by the demand-revealing process would desire constitutional restrictions that would limit the potential for overt redistribution. For instance, a proposal that a person's property be taken for a public purpose might be admissible only if he would be given reasonable compensation

Choices among Several Discrete Options

We now show how a demand-revealing process operates when there are more than two discrete options In table 2 there are three voters, indicated by numbers, and three options, indicated by letters The numbers in table 2 have been obtained by simply adding a third option, C, to the two shown in table 1, while leaving the differential the voter is willing to pay for a choice between A and B the same as it was in table 1 The difference between the numbers associated with any two options is interpreted as the amount of money the voter is willing to pay to have the one option with the higher number instead of the other Whether or not this is a legitimate interpretation will be discussed below

It may be noted that the rank-order preferences generate cyclic choices when majority rule is used To determine the collective choice by the demand-revealing process, we simply sum the columns and select the option with the highest total, which in this case is A There is no cycle, nor could there be, although a tie would be possible.

The tax for each voter is calculated from the lower portion of table 2 For the tax on voter 1, add up the sum of the other individuals' votes Option C would have been chosen with $70 The tax on voter 1 would be $70 minus $40, or $30, and voter 1 is better off by $20 ($50 – $30) than he would have been by abstaining Note that if he had understated his preferences enough to avoid being taxed, for instance, if he had reported that A only benefited him $25, C would have been selected and voter 1 would have been worse off than he was by correctly presenting his preferences In the case of voter 2, there is no tax because his vote does not change the outcome, and in the case of voter 3, there is a tax of $30 ($80 – $50), and he obtains a net benefit of $10 ($40 – $30) These taxes are fairly substantial, but that is because we have only a small number of voters With many voters, the probability is high that the total tax would be relatively miniscule if not zero

We next inquire whether the proposed method produces results that are "independent of irrelevant alternatives " If option C is dropped from the example in table 2, what difference would it make? Consider voter 1 first. He reports a differential value of $30 for A over B in table 2 If C were dropped from consideration, voter 1 would no longer have to offer $50 for A instead of C He would be richer, and he might spend some of his additional wealth to increase his offer with respect to A instead of B, say, from $30 to $32 Such wealth effects could conceivably change the result.

TABLE 2

AGGREGATING PREFERENCES FOR THREE OPTIONS (IN DOLLARS)

	DIFFERENTIAL VALUES OF OPTIONS				
VOTER	A	B	C	TAX	NET BENEFIT OF VOTING
1	50	20	0	30	20
2	0	60	20	0	0
3	40	0	50	30	10
Total	90	80	70		
		Total without Indicated Votes			
For 1· 2 + 3	40	60	70		
For 2 1 + 3	90	20	50		
For 3 1 + 2	50	80	20		

We do not, however, think that this is what is normally meant by a dependence on irrelevant alternatives Option C is relevant because its presence or absence affects the wealth of voter 1 If A owns a pizza den and is negotiating with B for its sale and C builds another pizza restaurant directly across the street, this will clearly affect the bargain between A and B However, we do not think that it would be proper to say that this was a situation which "lacked independence of irrelevant alternatives." In view of the general controversies surrounding this particular criterion of the Arrow theorem, however, we should like to simply discuss the wealth effect in the demand-revealing process rather than attempt to clear up the linguistic problem

To put the matter another way, when we insist that each voter arrange the options on a linear scale, so that the difference between the numbers on the scale for any pair of options represents what he is willing to pay to have one option instead of the other, we leave no room for wealth effects It may be that voter 1's true willingness to pay is $22 for B instead of C and $32 for A instead of B, but only $50 (rather than $54) for A instead of C, because if he has to pay $22 to get from C to B, he is poorer than if he starts at B With his lower wealth, it is not irrational for him to be willing to spend only $28 rather than $32, at that point, to get from B to A The linear scale does not permit voter 1 to report these wealth effects, so he compromises by reporting the values in table 2

It might be proposed that voters be asked to report their preferences among all pairs so that wealth effects could be taken into account in the decision process However, to do that would be to reintroduce a possibility of cycles. Consider the two voters shown in table 3.

Voter 1 has the preferences described earlier. Voter 2 has preferences in the opposite order, of approximately the same magnitude but with

TABLE 3

The Possible Cycle in Three Options (in Dollars)

	A against B		B against C		A against C	
1	32	0	22	0	50	0
2 ...	0	31	0	21	0	51

less nonlinearity from wealth effects When the preferences are summed, we find a collective choice for A over B and B over C, but also for C over A This problem of intransitivity might be resolved by applying some analytic device such as the "tournament matrix" described by Moon and Pullman (1970), but it is not clear how the Clarke tax would then be calculated. Furthermore, as long as there were cycles, there would be incentives for strategic misstatements Therefore, it may be best to require each voter to submit a linearized statement of his preferences, letting him make the necessary approximations there Then if he can guess which option would be selected without his vote, it will be in his interest to present comparisons with respect to that option truthfully.

A Simple Continuous Application

We now proceed with the specific case Clarke presented in his *Public Choice* article (1971). Assume there is some public good which can be purchased in any desired quantity For the purpose of graphic ease, we assume that it is sold in units which cost \$1, no matter how many are purchased, so that the line at \$1/unit on figure 1 represents the social cost schedule for purchasing different quantities The first stage in Clarke's process is to assign to each voter his share of the total cost Let us temporarily assume this share is assigned arbitrarily, and for the ith voter the share is the line shown at P_i. Later we will discuss how it is possible to approximate the Lindahl condition in the allocation of these shares

The voters are now asked to state their demand curves for the public good. Voter i's curve is shown as D_i. These curves are then summed vertically to get the aggregate demand (aggregate willingness to pay) curve *AD*. The point where the sum crosses the cost curve, that is, the \$1/unit line, is the efficient quantity of public good to purchase. This is, of course, the Samuelson equilibrium and has many fine properties, although not as many as the Lindahl equilibrium, toward which we shall move shortly.

How do we motivate voter i, and indeed all of the other voters, to correctly reveal their true demand curves? The answer is by telling each voter he will be subject to a Clarke tax, calculated as follows. When all the ballots are received, the tax for voter i will be calculated by summing

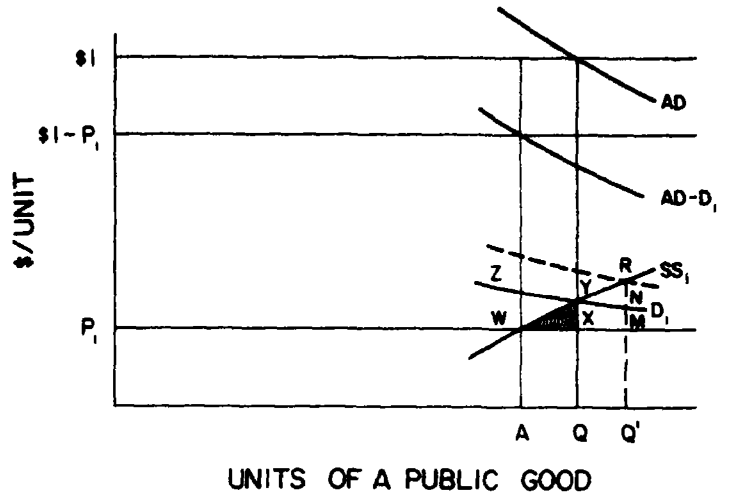


FIG. 1.—The tax on a person whose benefit exceeds his assigned tax share

(vertically) the demand curves of all of the voters other than i, generating the curve AD-D_i, and finding the intersection between that curve and the line $\$1$-$P_i$, which is the share of the tax cost that all voters other than i will pay. The intersection in figure 1 occurs at quantity A. This is the quantity of the public good that would be purchased if i reported a perfectly elastic (i.e., horizontal) demand schedule, identical to his cost share. Such a vote, offering to pay one's assigned share of whatever quantity others desire to purchase, is the analogue for continuous choices of abstaining in discrete choices. With voter i "abstaining," the revealed demand (of others) would intersect their share of cost at A, and i's payment would be the rectangle to the left of A and below P_i. In order to compute i's tax when he does not "abstain," we determine from the curve AD-D_i the amount of compensation that voter i would have to pay to keep all other voters indifferent to any change from quantity A. The required compensation per unit at any quantity is the difference at that quantity between the total cost and the aggregate willingness of others to pay. We call the schedule of such amounts, calculated as $\$1 - (AD - D_i)$, a synthetic supply schedule. It can be thought of as the net marginal social cost of supplying i with additional units of the public good after allowing a credit against the gross cost for the value of the good to others. This schedule is shown as line SS_i in figure 1. The schedule SS_i is a mirror image of AD-D_i. In this example, we assume that i has a higher willingness to pay than his cost share at A. This implies that the effect of including his demand is to increase the quantity. The intersection of the synthetic supply curve and i's demand is at the quantity Q, and that is the optimal amount of the public good, because that is also the quantity where AD intersects the \$1 line.

The amount that would have to be paid to individuals other than i in order to make them indifferent to the move from A to Q is represented by the area under SS_i, while the gain to voter i is the area under his demand curve. Voter i pays a composite tax which is the standard payment if he abstained plus the Clarke tax area under SS_i from A to Q. The total tax is equivalent to his assigned share of the cost of Q units of the public good (the rectangle to the left of line Q and below P_i) plus the shaded triangle WXY The sum of such rectangles for all voters is enough to pay the total cost of the public good, the shaded triangle and corresponding amounts for other voters must be wasted or given to nonvoters to keep all the incentives correct [2]

Suppose that voter i had misstated his demand curve in an effort to increase his net benefit. The benefit he has obtained from voting is the triangle WYZ Clearly, stating his demand as less than it actually is would reduce the size of his triangle On the other hand, if he stated his demand schedule as higher than it actually is, so that the quantity chosen would be, say, Q', his additional taxes would be $QQ'RY$ while his additional benefits would be only $QQ'NY$ He is best off correctly presenting his demand curve

It must be mentioned that there is a very slight conceptual problem in the specification of these demand curves Quantity demanded depends on income as well as price, and one determinant of a person's income is the Clarke tax he must pay Since one person's Clarke tax depends on the demand curves specified by others, each person could logically say that he could not specify his demand curve until all others had done so This is not a practical problem, however, because the Clarke tax, as we show below, is very small, and in most cases the uncertainty in the Clarke tax is very small, since it depends only on the elasticity of the aggregate willingness to pay of other voters, and in any event people can simply be directed to report demand curves that reflect their best guesses about their incomes

In figure 2, we assume that j, given his tax share, wants less than that amount of the public good which the other voters would choose. As in figure 1, D_j represents his true demand for the public good, and quantity A represents the amount which would be purchased if he chose to abstain, that is, the point where the sum of the demand curves of all the other voters intersects the sum of their shares of the tax price. Line SS_j in this case, as in figure 1, represents the rate of compensation per unit which it would be necessary to pay the other voters to compensate them for any

---

[2] One possibility for avoiding waste would be for pairs of communities to agree to exchange their collections of these excess revenues.

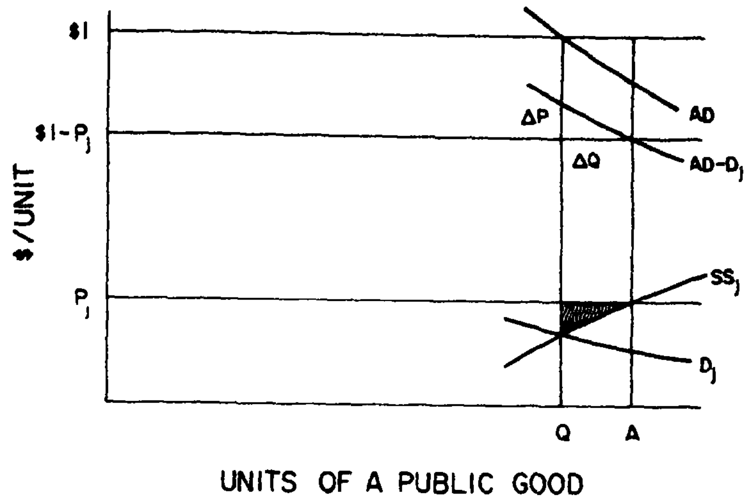


FIG 2 —The tax on a person whose benefit is less than his assigned tax share

change from point A in the amount of the public good. To the left of point A, SS_j may be interpreted as the rate of reduction in taxes for j that can be granted while reducing the quantity of the public good and reducing taxes for others by the full amount of the loss of income that they experience As in figure 1, this compensation is not actually going to be paid, but voter j will be taxed this amount

Once again, the point of intersection between j's demand curve and the synthetic supply curve represents the optimal quantity of the public good, Q, which is also the quantity at which AD intersects the $1/unit line In this case, Q is less than would be chosen if voter j abstained. Voter j then pays a tax which is equal to the rectangle to the left of line Q and below his tax share, plus the shaded triangle The rectangle is enough to pay his share of the cost of provision of the public good, the shaded triangle, once again, is wasted or given to nonvoters We will leave to the reader the demonstration that the correct presentation of his demand curve will maximize his welfare under these circumstances. It is essentially the same as the demonstration for figure 1

What is true for voters i and j is true for all voters. They are motivated by this peculiar tax procedure to present accurately their true demand curves The motivation, however, represented by triangle WYZ in figure 1, would normally be very small, about the same as the area of the shaded triangle.

To see how small the Clarke taxes would be, note that the shaded triangle in figure 2 is a mirror image of the one with sides labeled ΔP and ΔQ. If the elasticity of $AD - D_j$ is η, then $\Delta Q = \eta Q \Delta P/(1 - P_j)$, so that the area of the triangle is $1/2\eta Q(\Delta P)^2/(1 - P_j)$. The denominator approaches 1 as the number of voters increases, so that if η is on the order of magnitude of 2, then each voter's Clarke tax is roughly $Q(\Delta P)^2$. The values of ΔP would be related to the number of voters (N); it would be implausible for the average value of $(\Delta P)^2$ to be greater than $1/N^2$ Thus, the typical voter, whose share of the resource cost is Q/N, has a Clarke tax on the order of magnitude of $1/N$ times his resource cost, and the sum of all Clarke taxes is on the order of magnitude of one voter's taxes Thus, if the citizens of the United States were voting on the annual federal budget, the grand total of all the Clarke taxes charged would be in the neighborhood of $2,000, or about one-thousandth of a penny per person

As the triangles go to zero, the motivation for taking the trouble to present one's demand curve goes to zero The method is cheat-proof and generates the socially optimal quantity of the public good when every voter maximizes his self-interest, but when N is large the Downs paradox is present. voters have almost no incentive to vote

Since the excess revenues generated by the process are so very small, and certainly less than the administrative cost for any situation with more than a very small number of persons, the excess revenues deserve to be ignored This may be rather untidy, but it is normal in welfare economics to ignore the cost of reaching a decision If the Clarke tax is considered as part of the cost of making the decision, then it should be ignored. Contrarily, if it is not ignored, the cost of the processes of reaching decision rules by other processes should also be included We feel that our suggestion of simply wasting the extra revenue rather than searching for some complex budget-balancing process which might or might not achieve the same result is an important contribution It is also one of the reasons why it is so hard for welfare economists (ourselves included) to feel at home with the process.

So far, we have generated the Samuelson equilibrium; we now indicate how the Lindahl equilibrium may be approximated To this point, we have simply assigned the base share of the total expenditure for the individual in an arbitrary manner Suppose that, instead of assigning it arbitrarily, we appoint someone to do this, with the stipulation that from his pay we are going to subtract some multiple of the sum of the triangles for all of the voters.[3] The person assigning the fixed shares would be motivated to try to minimize the triangles. In the limit, if he were able to perfectly achieve his goal, there would be no triangles and no loss, we would have a perfect Lindahl equilibrium, with each voter paying for public goods according to his marginal evaluation.

---

[3] Probably the best way of selecting the "tax setter" would be to solicit bids. Precautions against bribery would, of course, be necessary.

It is unlikely that the official allocating the shares could do this perfectly, but he might be able to do quite well with advanced econometric methods. It should be emphasized, however, that there is one piece of information he cannot use in assigning the tax share of any individual: that individual's performance on previous choices. The voter, in making his choices on each individual decision, must not be able to offset against the optimality conditions for that particular choice the prospect of changing his base tax share in the future, because this would motivate him to misstate his demand curve.

As in all voting methods, there is a possibility for coalitions to distort the result. In particular, consider the strategy for a coalition of N persons whose equal benefits are greater than their equal tax shares. In figure 3, the demand schedule of voter i, D_i, is shown as a horizontal line because changes in the height would generally be negligible over the range of potential effects he and his coalition could have. The higher line, D_c, represents the demand that i would express taking account of the benefit of $D_i - P_i$ that each member of his coalition would receive for each unit increase in the quantity chosen. The distance from A to S is N times as far as the distance from A to Q, where the outcome would have been moved by an honest vote. With N persons in the coalition voting this way, the effect is to move the choice $N(N - 1)$ times the distance from A to Q, compared with honest voting. The gross benefit of the coalition activity to each member, in terms of benefits not paid for by his standard tax share, is $N(N - 1)(D_i - P_i)(Q - A)$, which is the area of the shaded rectangle. Each member's extra tax from coalition activity, apart from his standard tax share, is that portion of the shaded rectangle that is below the synthetic supply schedule. Thus, the net benefit of coalition activity to each member is a triangle like that in the upper left corner of the shaded rectangle, the area of which is proportional to $(N - 1)^2$ and to $(D_i - P_i)^2$. Thus, the benefit of forming coalitions varies with the square of the errors in tax shares and with the square of the number of members minus one. Voters whose tax shares overstate their benefits have a similar opportunity to form coalitions that multiply the understatements of their demands.

In this example, we have assumed that the only thing chosen is the unidimensional quantity of one public good. One of the convenient characteristics of the demand-revealing process is that it is not necessary to restrain voting to one issue at a time. A multidimensional public good or several public goods or public goods *plus* candidates can all be dealt with simultaneously. In general, it is much harder to organize coalitions in cases where the choice is not unidimensional. This is not to say that it is impossible. Still, we suspect that the demand-revealing process is rather less susceptible to coalition distortion than most voting methods.

The extension of the voting method to choices for more than one good

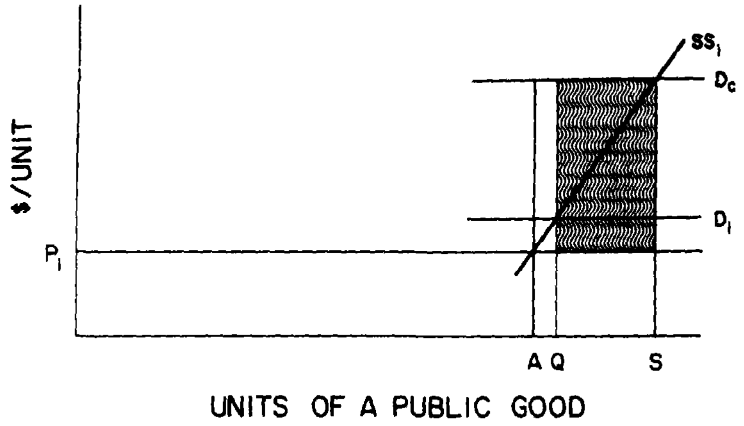


FIG 3 —The strategic calculation of a coalition

is straightforward if the chosen quantity of one public good has no impact on demands for other public goods. However, if the chosen quantity of some goods affects the demands for other goods, a simultaneous solution is needed One could ignore the interactions in the choice procedure and rely on individuals to make estimates of the quantities of other goods that would be chosen in reporting their demand schedules, but any misestimates by voters would lead to unnecessary inefficiencies

At a conceptual level, one could ask all voters to report their marginal valuation schedules for each good at every combination of quantities of other public goods, although the data problem if this were really attempted would be unmanageable If it were not impossible to obtain and operate on the data, the identification of an equilibrium where the appropriate marginal conditions were all satisfied simultaneously would be essentially no different from calculating a competitive equilibrium for private goods Groves and Ledyard (1975) developed the theoretical foundations of such a system in detail

In later publications,[4] we propose to apply the process to a number of other problems such as income redistribution, badly behaved demand curves, and use as a welfare indicator We will also discuss its practical application in realistic government structure The purpose of this paper, however, has been to explain the system and to demonstrate that it solves a number of problems previously thought to be unsolvable The process does not violate the Arrow theorem, but it avoids the problems of the Arrow theorem by not meeting Arrow's assumptions However, it seems to us that, if the Arrow theorem is considered as a result that suggests that a good voting process cannot be devised, then the real problem raised by Arrow is solved by this process.

---

[4] Mimeographed preliminary drafts are available on request.

References

Bowen, H. R. "The Interpretation of Voting in the Allocation of Economic Resources." *Q J.E.* 58 (November 1943) 32–42

Clarke, E. H. "Multipart Pricing of Public Goods" *Public Choice* 11 (Fall 1971). 17–33.

———. "Multipart Pricing of Public Goods An Example" In *Public Prices for Public Products*, edited by S. Mushkin. Washington Urban Inst, 1972.

Downs, A *An Economic Theory of Democracy*. New York Harper & Row, 1957.

Dreze, J H, and de la Vallee Poussin, D "A Tâtonnement Process for Public Goods." *Rev. Econ Studies* 38 (April 1971) 133–50

Groves, T "Incentives in Teams" *Econometrica* 41 (July 1973) 617–33

Groves, T., and Ledyard, J "Optimal Allocation of Public Goods A Solution to the 'Free Rider Problem'" Discussion paper 144 Center Math. Studies Econ and Management Sci, Northwestern Univ, May 1975

Groves, T, and Loeb, M "Incentives and Public Inputs" *J Public Econ.* 4 (August 1975) 211–26

Moon, J W, and Pullman, N J "On Generalized Tournament Matrices" *SIAM Rev* 12 (July 1970)· 389–94.

Thompson, E. "A Pareto Optimal Group Decision Process." *Papers on Non-Market Decision Making* 1 (1965) 133–40

Tideman, T N "The Efficient Provision of Public Goods" In *Public Prices for Public Products*, edited by S Mushkin Washington Urban Inst, 1972

Vickrey, W. "Counterspeculation, Auctions, and Competitive Sealed Tenders" *J. Finance* 16 (May 1961)· 8–37.

Expectations and Exchange Rate Dynamics

Rudiger Dornbusch

Massachusetts Institute of Technology

The paper develops a theory of exchange rate movements under perfect capital mobility, a slow adjustment of goods markets relative to asset markets, and consistent expectations. The perfect foresight path is derived and it is shown that along that path a monetary expansion causes the exchange rate to depreciate. An initial overshooting of exchange rates is shown to derive from the differential adjustment speed of markets. The magnitude and persistence of the overshooting is developed in terms of the structural parameters of the model. To the extent that output responds to a monetary expansion in the short run, this acts as a dampening effect on exchange depreciation and may, in fact, lead to an increase in interest rates

I. Introduction

The paper develops a simple macroeconomic framework for the study of exchange rate movements The purpose is to develop a theory that is suggestive of the observed large fluctuations in exchange rates while at the same time establishing that such exchange rate movements are consistent with rational expectations formation In developing a formal model we draw on the role of asset markets, capital mobility, and expectations that have been emphasized in recent literature.[1] We draw, too, on the fact

I am indebted to Stanley Black, Franco Modigliani, and Edward Tower who provided the stimulus for this paper In revising various drafts I have had the benefit of many comments I wish, in particular, to acknowledge the helpful suggestions I have received from Wilfred Ethier, Stanley Fischer, Jacob Frenkel, and the thoughtful remarks of two anonymous referees. Financial support was provided by a grant from the Ford Foundation

[1] For recent work on flexible exchange rates that shares some of the present emphasis, see Black (1973, 1975), Henderson (1975), Niehans (1975), Dornbusch (1976*a*, 1976*b*), Frenkel (1976), Kouri (1976), and Mussa (1976). The classics remain Fleming (1962) and Mundell (1964, 1968).

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved.

of differential adjustment speeds in goods and asset markets. In fact, the dynamic aspects of exchange rate determination in this model arise from the assumption that exchange rates and asset markets adjust fast relative to goods markets

The adjustment process to a monetary expansion in this framework serves to identify several features that are suggestive of recent currency experience In the short run, a monetary expansion is shown to induce an immediate depreciation in the exchange rate and accounts therefore for fluctuations in the exchange rate and the terms of trade Second, during the adjustment process, rising prices may be accompanied by an appreciating exchange rate so that the trend behavior of exchange rates stands potentially in strong contrast with the cyclical behavior of exchange rates and prices The third aspect of the adjustment process is a direct effect of the exchange rate on domestic inflation In this context the exchange rate is identified as a critical channel for the transmission of monetary policy to aggregate demand for domestic output

The effect of monetary policy on interest rates and exchange rates is significantly affected by the behavior of real output. If real output is fixed, a monetary expansion will, in the short run, lower interest rates and cause the exchange rate to overshoot its long-run depreciation If output, on the contrary, responds to aggregrate demand, the exchange rate and interest rate changes will be dampened While the exchange rate will still depreciate, it may no longer overshoot, and interest rates may actually rise

In Part II we develop a formal model in terms of explicit functional forms That development allows us to derive an analytical solution for the time path of variables and, in Part III, for the expectations that generate the perfect foresight path In Part IV, the model is used to investigate the effects of a monetary disturbance While the major part of the analysis is developed for the case of fixed output, an extension to variable output is introduced in Part V

II. The Model

We will assume a country that is small in the world capital market so that it faces a given interest rate Capital mobility will ensure the equalization of expected net yields so that the domestic interest rate, less the expected rate of depreciation, will equal the world rate. In the goods market we will assume that the world price of imports is given. Domestic output is an imperfect substitute for imports, and aggregate demand for domestic goods, therefore, will determine their absolute and relative price.

A Capital Mobility and Expectations

Assets denominated in terms of domestic and foreign currency are assumed to be perfect substitutes given a proper premium to offset anticipated exchange rate changes. Accordingly, if the domestic currency is expected to depreciate, interest rates on assets denominated in terms of domestic currency will exceed those abroad by the expected rate of depreciation. That relationship is expressed in (1) where r is the domestic interest rate, r^* is the given world rate of interest, and x is the expected rate of depreciation of the domestic currency, or the expected rate of increase of the domestic currency price of foreign exchange

$$r = r^* + x \tag{1}$$

Equation (1) is a representation of perfect capital mobility, and it is assumed that incipient capital flows will ensure that (1) holds at all times

Consider next expectations formation Here we distinguish between the long-run exchange rate, to which the economy will ultimately converge, and the current exchange rate Denoting the logarithms of the current and long-run rate by e and $\bar{e}$, respectively, we assume that

$$x = \theta(\bar{e} - e) \tag{2}$$

Equation (2) states that the expected rate of depreciation of the spot rate is proportional to the discrepancy between the long-run rate and the current spot rate The coefficient of adjustment θ is for the present taken as a parameter The long-run exchange rate is assumed known, and an expression for it will be developed below We note further that, while expectations formation according to (2) may appear ad hoc, it will actually be consistent with perfect foresight, as shown in Part III

B The Money Market

The domestic interest rate is determined by the condition of equilibrium in the domestic money market The demand for real money balances is assumed to depend on the domestic interest rate and real income and will, in equilibrium, equal the real money supply Assuming a conventional demand for money, the log of which is linear in the log of real income and in interest rates, we have[2]

$$-\lambda r + \phi y = m - p, \tag{3}$$

where m, p, and y denote the logs of the nominal quantity of money, the price level, and real income For the remainder of this part we will take the nominal quantity of money and the level of real income as given

Combining (1), (2), and (3) will give us a relationship between the spot exchange rate, the price level, and the long-run exchange rate, given that the money market clears and net asset yields are equalized:

$$p - m = -\phi y + \lambda r^* + \lambda\theta(\bar{e} - e). \tag{4}$$

---

[2] Equation (3) is obtained by taking the logarithm of the money market equilibrium condition $M/P = Y^{\phi} \exp(-\lambda r)$.

Equation (4) can be simplified by noting that with a stationary money supply long-run equilibrium will imply equality between interest rates, because current and expected exchange rates are equal. This implies that the long-run equilibrium price level, $\bar{p}$, will equal

$$\bar{p} = m + (\lambda r^* - \phi y). \tag{5}$$

Substituting (5) in (4) gives us a relationship between the exchange rate and the price level:[3]

$$e = \bar{e} - (1/\lambda\theta)(p - \bar{p}) \tag{6}$$

Equation (6) is one of the key equations of the model. For given long-run values of exchange rates and prices, it serves to determine the current spot price of foreign exchange as a function of the current level of prices Given the level of prices, we have a domestic interest rate and an interest differential Given the long-run exchange rate, there is a unique level of the spot rate such that the expected appreciation, or depreciation, matches the interest differential An increase in the price level, because it raises interest rates, gives rise to an incipient capital inflow that will appreciate the spot rate to the point where the anticipated depreciation exactly offsets the increase in domestic interest rates.

C. *The Goods Market*

The demand for domestic output depends on the relative price of domestic goods, $e - p$, interest rates, and real income The demand function is assumed to have the form

$$\ln D = u + \delta(e - p) + \gamma y - \sigma r, \tag{7}$$

where D denotes the demand for domestic output and where u is a shift parameter[4] From (7) we note that a decrease in the relative price of domestic goods raises demand, as does an increase in income or a reduction in interest rates. The rate of increase in the price of domestic goods, $\dot{p}$, is described in (8) as proportional to an excess demand measure

$$\dot{p} = \pi \ln (D/Y) = \pi[u + \delta(e - p) + (\gamma - 1)y - \sigma r] \tag{8}$$

---

[3] In (3) we assumed that the appropriate deflator for money balance is the price of domestic output An alternative is provided by a deflator that is a weighted average of domestic and import prices. In such a formulation the "price level," q, could be written as $q = \alpha p + (1 - \alpha)e$, where α and $(1 - \alpha)$ are the expenditure shares of domestic goods and imports. With such a formulation (6) would be amended to the following equation: $e = \bar{e} - \beta(p - \bar{p})$, where $\beta = \alpha/[\lambda\theta + (1 - \alpha)]$. None of the qualitative results described below would be affected by this extension.

[4] The complete relative price argument in (7) is $(e + p^* - p)$ where p^* is the logarithm of the foreign price level. Setting the foreign price level equal to unity implies that $p^* = 0$.

We note that the long-run equilibrium exchange rate implied by (8) is[5]

$$\bar{e} = \bar{p} + (1/\delta)[\sigma r^* + (1 - \gamma)y - u], \tag{9}$$

where $\bar{p}$ is defined in (5) From (9) it is apparent that the long-run exchange rate depends with the conventional homogeneity properties on monetary variables, but obviously on real variables, too

The price equation in (8) can be simplified by using the definition of the long-run rate in (9) and the fact that interest differences equal expected depreciation, $r - r^* = \theta(\bar{e} - e)$, to become[6]

$$\dot{p} = -\pi[(\delta + \sigma\theta)/\theta\lambda + \delta](p - \bar{p}) = -v(p - \bar{p}), \tag{10}$$

where

$$v \equiv \pi[(\delta + \sigma\theta)/\theta\lambda + \delta] \tag{11}$$

The price adjustment equation in (10) can be solved to yield

$$p(t) = \bar{p} + (p_0 - \bar{p}) \exp(-vt), \tag{12}$$

which shows that the price of domestic output will converge to its long-run level at a rate determined by (11) Substitution of (12) in (6) gives the time path of the exchange rate

$$\begin{aligned} e(t) &= \bar{e} - (1/\lambda\theta)(p_0 - \bar{p}) \exp(-vt) \\ &= \bar{e} + (e_0 - \bar{e}) \exp(-vt) \end{aligned} \tag{13}$$

From (13) the exchange rate will likewise converge to its long-run level The rate will appreciate if prices are initially below their long-run level and, conversely, if prices initially exceed their long-run level

D Equilibrium Exchange Rates

The adjustment process of the economy can be described with the help of figure 1 At every point in time the money market clears and expected yields are arbitraged This implies a relationship between prices and the spot exchange rate shown in (6) and reflected in the *QQ* schedule in figure 1. The positively sloped schedule $\dot{p} = 0$ shows combinations of price levels and exchange rates for which the goods market and money market are in equilibrium [7] Points above and to the left of that schedule

---

[5] Equation (9) is obtained by setting $\dot{p} = 0$ and $r = r^*$ as is appropriate for the long run where markets clear and exchange rates are constant

[6] In (8) aggregate demand depends on the nominal interest rate. An alternative formulation allows aggregate demand to depend on the real interest rate, $r - \dot{p}$. Such a formulation requires that we substitute $\rho \equiv \pi/(1 - \sigma\pi) > 0$ in place of π in (11) and the equations below The restriction that $\rho > 0$ is required for stability.

[7] The $\dot{p} = 0$ schedule represents combined goods- and money-market equilibrium. Setting $\dot{p} = 0$ in (8) and substituting for the domestic interest rate from (3) yields the equation of the goods-market equilibrium schedule

$$p = [\delta\lambda/(\delta\lambda + \sigma)]e + [\sigma/(\delta\lambda + \sigma)]m + [\lambda/(\delta\lambda + \sigma)][u + (1 - \gamma)y - \phi\sigma y/\lambda].$$

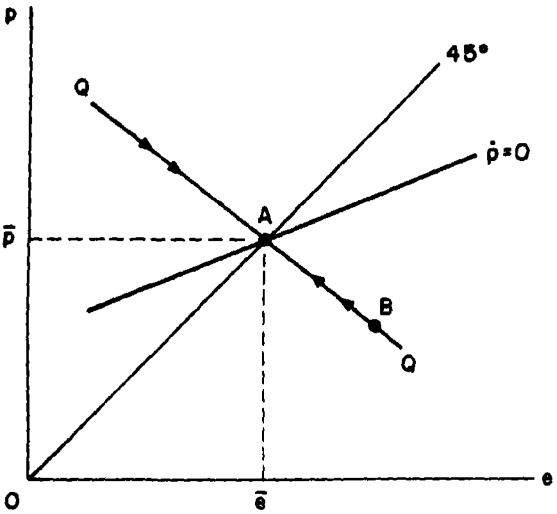


FIG 1

correspond to an excess supply of goods and falling prices Conversely, points to the right and below the schedule correspond to an excess demand The $\dot{p} = 0$ schedule is positively sloped and flatter than a 45° line for the following reason [8] An increase in the exchange rate creates an excess demand for domestic goods by lowering their relative price To restore equilibrium, domestic prices will have to increase, though proportionately less, since an increase in domestic prices affects aggregate demand, both via the relative price effect and via higher interest rates.

For any given price level the exchange rate adjusts instantaneously to clear the asset market Accordingly, we are continuously on the *QQ* schedule with money-market equilibrium and international arbitrage of net expected yields. Goods-market equilibrium, to the contrary, is only achieved in the long run Conditions in the goods market, however, are critical in moving the economy to the long-run equilibrium by inducing rising or falling prices. Specifically, an initial position such as point *B*, with a price level below the long-run level and, correspondingly, an exchange rate in excess of the long-run equilibrium, implies an excess demand for goods because domestic output commands a low relative price and because the interest rate is low. Accordingly, prices will be rising, thereby inducing over time a reduction in excess demand. The path of rising prices is accompanied by an appreciation of the exchange rate. As interest rates rise, as a consequence of declining real balances, the spot rate will approach the long-run rate. Once the long-run equilibrium at point A is attained, interest rates are equal internationally, the goods market clear, prices are constant, and expected exchange rate changes are zero.

---

[8] The 45° line in fig 1 is drawn through the origin on the assumption that, by appropriate choice of units, the prices of both goods are initially equal.

III. Consistent Expectations

So far we have placed no restrictions on the formation of expectations other than the assumption that the expected rate of depreciation, as shown in (2) is proportional to the discrepancy between the long-run and the current exchange rate From (12) and (13) we note that the rate at which prices and the exchange rate converge to equilibrium is given by v From (11) it is apparent that the rate of convergence is a function of the expectations coefficient, θ

Clearly, for the expectations formation process in (2) to correctly predict the actual path of exchange rates it must be true that $\theta = v$. Accordingly, the expectations coefficient, θ, that corresponds to perfect foresight, or, equivalently, that is consistent with the model is given by the solution to the equation

$$\theta = v \equiv \pi[(\delta + \sigma\theta)/\theta\lambda + \delta] \qquad (14)$$

The consistent expectations coefficient, $\tilde{\theta}$, obtained as the solution to (14), is a function of the structural parameters of the economy[9]

$$\tilde{\theta}(\lambda, \delta, \sigma, \pi) = \pi(\sigma/\lambda + \delta)/2 + [\pi^2(\sigma/\lambda + \delta)^2/4 + \pi\delta/\lambda]^{1/2}. \qquad (15)$$

Equation (15) gives the rate at which the economy will converge to long-run equilibrium along the perfect foresight path If expectations are formed according to (2) and (15), exchange rate predictions will actually be borne out[10] The characteristics of the perfect foresight path are that the economy will converge faster the lower the interest response of money demand and the higher the interest response of goods demand and the price elasticity of demand for domestic output The reason is simply that with a low interest response a given change in real balances will give rise to a large change in interest rates which, in combination with a high interest response of goods demand, will give rise to a large excess demand and therefore inflationary impact. Similarly, a large price elasticity serves to translate an exchange rate change into a large excess demand and, therefore, serves to speed up the adjustment process.

---

[9] In (16) we have taken the positive and therefore stable root of the quadratic equation implied by (14)

[10] Perhaps a remark about the perfect foresight path is in order here. Why should that path command our interest rather than being a mere *curiosum*? The reason is that it is the only expectational assumption that is not arbitrary (given the model) and that does not involve persistent prediction errors The perfect foresight path is, obviously, the deterministic equivalent of rational expectations

IV. The Effects of a Monetary Expansion

In this part we will study the adjustment process to a monetary expansion. The analysis serves to derive substantive results but also to highlight the manner in which expectations about the future path of the economy affect the current level of the exchange rate. This link is embodied in consistent expectations and makes the impact effect of a monetary disturbance depend on the entire structure of the economy.

In figure 2 we show the economy in initial full equilibrium at point A, with a long-run price level $\bar{p}$ and a corresponding long-run exchange rate $\bar{e}$ where the level of prices is determined, according to (5), by the nominal quantity of money, real income, and the interest rate. The long-run exchange rate by (9) will depend on the level of domestic prices and characteristics of the demand for domestic goods. The asset-market equilibrium schedule QQ that combines monetary equilibrium and arbitrage of net expected yields is drawn for the initial nominal quantity of money.

An increase in the nominal quantity of money that is expected to persist will cause a goods and asset market disequilibrium at the initial exchange rate and price. To maintain asset-market equilibrium, the increased quantity of money would have to be matched by higher prices and/or a depreciation in the exchange rate. The asset-market equilibrium schedule will shift out to $Q'Q'$, a shift that is (proportionately) equal to the increase in the nominal quantity of money.

It is immediately obvious that the new long-run equilibrium is at point C, where both goods and asset markets clear and exchange rate and price changes exactly reflect the increase in money.[11] This long-run homogeneity result is not surprising, since there is no source of money illusion or long-run price rigidity in the system.

Consider next the adjustment process. At the initial level of prices, the monetary expansion reduces interest rates and leads to the anticipation of a depreciation in the long run and, therefore, at the current exchange rate, to the expectation of a depreciating exchange rate. Both factors serve to reduce the attractiveness of domestic assets, lead to an incipient capital outflow, and thus cause the spot rate to depreciate. The extent of that depreciation has to be sufficient to give rise to the anticipation of appreciation at just sufficient a rate to offset the reduced domestic interest rate. The impact effect of a monetary expansion is, therefore, to induce an immediate depreciation in the spot rate and one that exceeds the long-run depreciation, since only under these circumstances will the public anticipate an appreciating exchange rate and thus be compensated for the reduced interest on domestic assets. This is shown in figure 2 by the move from point A to the short-run equilibrium at point B.

---

[11] We have not drawn in fig. 2, the $\dot{p} = 0$ schedule. It is apparent, however, from the homogeneity properties of the model that the $\dot{p} = 0$ schedule will pass through point C.

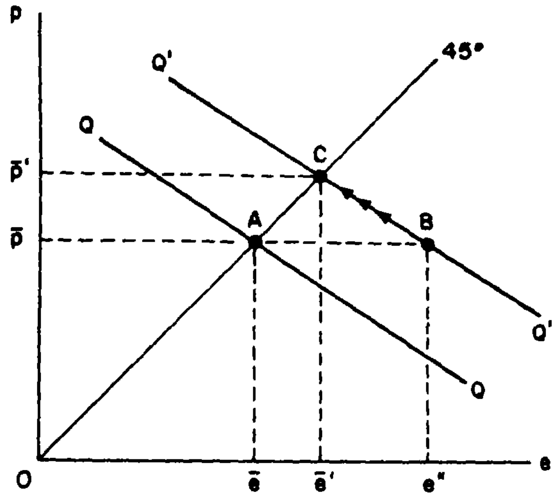


FIG 2

From (4), noting that $d\bar{e} = dm = d\bar{p}$, we obtain a formal expression for the impact effect of a monetary expansion on the spot exchange rate:

$$de/dm = 1 + 1/\lambda\theta \qquad (16)$$

Equation (16) confirms that in the short run the exchange rate will overshoot. The extent of the overshooting will depend on the interest response of money demand and the expectations coefficient.

A high interest response of money demand will serve to dampen the overshooting because it implies that a given expansion in the (real) quantity of money will only induce a small reduction in the interest rate. A small reduction in the interest rate in turn requires only a small expectation of appreciation to offset it and therefore, given the coefficient of expectations and the long-run rate, only a small depreciation of the spot rate (in excess of the long-run rate) to generate that expectation. A similar interpretation applies to the coefficient of expectations in (16).

It is quite obvious from the preceding explanation that the short-term effects of a monetary expansion, in this model, are entirely dominated by asset markets and, more specifically, by capital mobility and expectations. This feature places in sharp relief the assumption that asset markets and exchange rates adjust fast relative to the goods market and the price of domestic output. It is under these circumstances that a change in the nominal quantity of money is, in fact, a change in the real quantity of money, and the spot rate adjustment serves to achieve equilibrium in the asset markets by creating the expectation of appreciation of just sufficient an extent to balance the reduced interest rate on domestic assets.

The interpretation of (16) has not so far used the restriction that expectations be rational. That restriction is introduced by substituting (15) in (16) to obtain

$$de/dm = 1 + 1/\lambda\tilde{\theta} = 1 + \frac{1}{\pi(\sigma + \delta\lambda)/2 + [\pi^2(\sigma + \delta\lambda)^2/4 + \pi\delta\lambda]^{1/2}} \tag{17}$$

Equation (17) has two implications that cannot be derived from (16) The first is that with an interest response of money demand that approaches zero the initial depreciation remains finite and, in fact, approaches $de/dm = 1 + 1/\pi\sigma$. This result reflects the fact that, for the large interest rate changes that would result in these circumstances, the subsequent path of prices and the exchange rate is governed by the effect of interest rates on aggregate demand

A second implication of (17) is the fact that the short-run overshooting of the exchange rate is inversely related to the speed of adjustment of the system, $\tilde{\theta}$ That fact is particularly obvious for the case where the speed of adjustment of prices, π, becomes infinite and where, accordingly, the economy jumps instantaneously to the new long-run equilibrium at point C[12] More generally, those factors that serve to speed up the adjustment process, in particular high interest rate responsiveness of money demand, or aggregate spending, or high price elasticities, will therefore serve to dampen the impact effect of a monetary expansion on the exchange rate This effect relies entirely on expectations about the subsequent path of the economy, rather than on *current* interaction between goods and asset markets

Consider next the adjustment process from the short-run market equilibrium at point B to long-run equilibrium at point C We note from figure 2 that at point B there is an excess demand for goods That excess demand arises both from the decline in domestic interest rates and from the depreciation in the exchange rate that lowers the *relative* price of domestic goods Each factor by itself is sufficient to account for this excess demand and, in fact, they constitute independent channels through which monetary changes affect demand for domestic output

The exchange rate channel has been identified by Fleming and Mundell as an important avenue for monetary policy to act on aggregate demand.[13] In the present context the depreciation of the spot rate that is induced by the conditions of asset-market equilibrium serves to reduce the relative price of domestic goods and thereby to raise aggregate demand and give rise to inflationary pressure as opposed to an increase in output The importance of this channel is larger, the higher the price elasticity of demand relative to the interest response of aggregate spending.

---

[12] The slope of the QQ schedule is $dp/de = -\tilde{\theta}\lambda$, and the schedule becomes vertical as $\tilde{\theta}$ approaches infinity

[13] In the Mundell–Fleming model with prices and interest rates fixed, the depreciation by worsening the terms of trade creates the necessary increase in aggregate demand to support the higher level of income required by monetary equilibrium (for a further discussion see Niehans [1975] and Dornbusch [1976])

The lower interest rates and a lower relative price of domestic goods, that are characteristics of the impact effect, will cause domestic prices to rise and therefore be reflected in falling real money balances, rising interest rates, and an appreciating exchange rate The adjustment process of rising prices over time restores the economy to the initial real equilibrium An important feature of that adjustment process is the fact that rising prices are accompanied by an appreciating exchange rate In terms of figure 2, this is described by the move along $Q'Q'$ from B to C This result is due to the fact that rising prices cause the real money supply to be falling and interest rates to be rising The rising interest rate, in turn, gives rise to an incipient capital inflow that appreciates the exchange rate at the same rate as interest rates are rising and thus maintains expected net yields in line The model therefore confirms the link between interest rates and exchange rates that is emphasized in popular interpretations of foreign exchange events The observation is correct, in the present circumstances, because rising interest rates are accompanied by the expectation of an appreciating exchange rate

In summarizing this part we note that the ultimate effect of a monetary expansion is an equiproportionate increase in prices and the exchange rate In the short run, however, the monetary expansion does exert real effects on interest rates, the terms of trade, and aggregate demand The details of the adjustment process will depend on the economic structure In particular, terms of trade changes will be both larger and more persistent the lower the speed of adjustment, $\tilde{\theta}$

A key role in this analysis is played by the sluggish adjustment of prices as compared with asset markets There is no very persuasive theoretical support for the slow adjustment of goods markets, but the facts clearly point in this direction While the differential adjustment speed lacks theoretical backing, it implies, nevertheless, a behavior of exchange rates that is suggestive of recent experience[14]

V. Short-Run Adjustment in Output

So far we have assumed that output is fixed at the full-employment level, $\bar{y}$. In the present part, the analysis is extended to allow for short-run adjustments in output in response to changes in aggregate demand

---

[14] An extension of this paper would draw in an explicit manner on stochastic elements to provide a rationale for the short-run stickiness of prices At the same time, such an extension would have interesting implications for the manner in which expectations are formed Exchange rate determination in a stochastic setting has been studied by Black (1973), Kouri (1975), and Mussa (1976). Fischer (1976) has used a stochastic framework to evaluate fixed versus flexible exchange rate systems

Therefore, we replace equation (8) by an equilibrium condition in the domestic goods market,

$$y = \ln D \equiv u + \delta(e - p) + \gamma y - \sigma r, \tag{18}$$

where y is the log of the actual level of output that in the short run is demand determined In addition to (18), we require a price adjustment equation which is shown in (19):

$$\dot{p} = \pi(y - \bar{y}) \tag{19}$$

According to (19) the rate of inflation is proportional to the discrepancy between actual and full employment, or "potential" output, $\bar{y}$ This price adjustment equation is a combination of a relationship between wage and price inflation, a relation between wage inflation and unemployment as in a Phillips curve, and a relation between unemployment and the departure from potential output, $y - \bar{y}$, as described by Okun's law

It is shown in the Appendix below that the extension that incorporates (18) and (19) in place of (8) leaves most of the analysis of adjustments to a monetary increase unchanged In particular, the price adjustment will continue to be exponential although the speed of adjustment will depend now also on the income elasticities of demand for domestic goods and real balances, γ and ϕ

In the present framework it continues to be true that in the short run an increase in the nominal quantity of money is an increase in the real quantity of money Accordingly, a monetary expansion has the conventional effect of increasing in the short run the level of output and inducing inflation. Since the inflation that is induced by the expansion in real output serves to raise over time the price level, real balances will decline back to their initial level until in the long run the expansion in money is fully matched by increased prices and output has returned to the full-employment level

The impact effect of a monetary expansion on exchange rates and interest rates may, however, differ significantly from the analysis in Part IV The new possibility that arises from the expansion in output in the short run is that the exchange rate depreciation will fall short of the monetary expansion rather than exceed it as in (16). That possibility arises because, in the short run, the income expansion raises money demand and may do so sufficiently to actually increase interest rates If the output expansion were sufficiently strong to raise interest rates, equalization of net yields internationally would require the expectation of a depreciation and therefore a spot rate that falls short of the long-run equilibrium rate. Since the long-run equilibrium rate increases in the same proportion as the nominal quantity of money, it follows that the spot rate would increase less than the quantity of money. The condition that gives rise to this case is

$$1 - \phi\delta/(1 - \gamma) < 0. \tag{20}$$

The term $\delta/(1 - \gamma)$ is the elasticity of equilibrium output with respect to the exchange rate That term multiplied by the income elasticity of demand gives the increase in money demand due to a depreciation in the spot rate Accordingly, (20) tells us whether at constant interest rates, and allowing the exchange rate to depreciate in the same proportion as the increase in money, we have an excess demand or supply of money and, accordingly, an increase or decrease in interest rates The possibility of an excess demand and therefore an increase in interest rates is associated with a high income elasticity of money demand, high price elasticity, and a high income elasticity of demand for domestic goods

The time path of exchange rates and the interest rate therefore depends on income and price elasticities, and the short-run overshooting of exchange rates is no longer a necessary feature of the adjustment process. In fact, if in the short run the interest rate rises and the exchange rate therefore depreciates less than proportionately to the increase in money, the adjustment process will be one of rising prices and a depreciating exchange rate In this event, therefore, terms of trade fluctuations will be dampened as compared with the case described earlier where the exchange rate overshooting introduces large terms of trade variations in the adjustment process

The analysis of a monetary expansion in this part confirms once more the Mundell-Fleming result that under conditions of capital mobility and flexible rates a small country can conduct, in the short run, an effective monetary policy More important, the exchange rate proves a critical channel for the transmission of monetary changes to an increase in aggregate demand and output. That channel may, in fact, prove to be the only channel since, as was shown above, the interest rate may actually rise in the transition Unlike in the Mundell-Fleming world, extension of the analysis to the long run shows that the effects of a monetary expansion are only transitory, since the inflation that is induced by the output expansion serves to reduce real balances and thereby return interest rates, relative prices, and real income to their initial level

The possibility of short-run output adjustment has been shown to dampen exchange rate movements and possibly reverse the interest rate effects of a monetary expansion It is appropriate, therefore, to ask which of the assumption, fixed or variable output, is a more relevant characteristic of the adjustment process. The answer no doubt is that the fixed output adjustment is a suitable characterization of the very short run. In the very short run we would not expect output to adjust instantaneously to meet an increase in aggregate demand and, accordingly, the adjustment will be primarily confined to the asset markets and will be characterized by a decline in interest rates and overshooting of exchange rates. In the intermediate run, on the contrary, the present analysis gains relevance, since here we would expect an adjustment of both output and prices in response to increased aggregate demand On balance, therefore, the fixed output case retains relevance, and particularly so if output adjusts sluggishly to changes in aggregate demand

Appendix

This Appendix extends the model to include short-run supply responses For that purpose we replace the price adjustment equation in (8) by a goods-market equilibrium condition (A1) and a price equation (A2)

$$y = \mu[u + \delta(e - p) - \sigma r], \mu \equiv 1/(1 - \gamma) > 0, \tag{A1}$$

$$\dot{p} = \pi(y - \bar{y}), \tag{A2}$$

where $\bar{y}$ denotes the full-employment level of output and where the price adjustment equation can be thought of as arising from a Phillips-curve relation between wage inflation and unemployment combined with an Okun's-law relation between the deviation from potential output, $y - \bar{y}$, and unemployment [15]

The specification of the money market and exchange rate expectations remains unchanged, and equation (4) that represents these relations is repeated here for convenience

$$p - m + \phi y = \lambda r^* + \theta\lambda(\bar{e} - e) \tag{A3}$$

Noting that in long-run equilibrium we have $y = \bar{y}$ and $r = r^*$, we obtain from (A1) the long-run goods-market relationship

$$\bar{y} = \mu[u + \delta(\bar{e} - \bar{p}) - \sigma r^*], \tag{A4}$$

and subtracting (A4) from (A1) we obtain the goods-market equilibrium condition expressed in terms of deviations from long-run equilibrium,

$$y - \bar{y} = \mu(\delta + \sigma\theta)(e - \bar{e}) + \mu\delta(\bar{p} - p), \tag{A5}$$

where we have used the fact that $r^* - r = \theta(e - \bar{e})$.

Next we proceed in a similar manner for the money market and rewrite the equilibrium condition as

$$\phi(y - \bar{y}) + (p - \bar{p}) = \lambda\theta(\bar{e} - e). \tag{A6}$$

Equations (A5) and (A6) can be simultaneously solved to yield the spot exchange rate and the level of output as a function of the existing price level These solutions are, respectively,

$$y - \bar{y} = -w(p - \bar{p}), \tag{A7}$$

where

$$w \equiv [\mu(\delta + \theta\sigma) + \mu\delta\theta\lambda]/\Delta; \qquad \Delta \equiv \phi\mu(\delta + \theta\sigma) + \theta\lambda,$$

and

$$e - \bar{e} = -[(1 - \phi\mu\delta)/\Delta](p - \bar{p}). \tag{A8}$$

---

[15] To deal with steady-state inflation we would have to add in (A2) the long-run rate of inflation which is given by the rate of monetary growth, which in the present treatment is assumed equal to zero.

Substitution of (A7) in (A2) yields the equilibrium rate of inflation as a function of the price level:

$$\dot{p} = -\pi w(p - \bar{p}). \tag{A9}$$

Following the procedure in Part III, rational expectations require that the expectations coefficient, θ, equal the rate at which exchange rates actually adjust, πw.

$$\theta = \pi w, \tag{A10}$$

which can be solved for the rational expectations coefficient of adjustment, $\tilde{\theta}$.

Consider next the impact effect of a monetary expansion. Remembering that in the long run an increase in money causes an equiproportionate increase in prices and the exchange rate, we have $d\bar{e} = d\bar{p} = dm$. Therefore, from (A8) we obtain the impact effect of a monetary expansion on the exchange rate as

$$de/dm = 1 + (1 - \phi\mu\delta)/\Delta > 0 \tag{A11}$$

Whether the exchange rate increases more or less proportionately than the nominal quantity of money depends on the condition

$$1 - \phi\mu\delta \gtrless 0, \tag{A12}$$

which determines, too, whether the interest rate declines or increases.

By (A7) the impact effect on real output is unambiguously positive and equal to $dy/dm = w$. The increase in the rate of inflation is given by $d\dot{p}/dm = \pi w$.

Since from (A9) the inflation rate converges monotonically to the long-run level, we know that output declines monotonically back toward the level of full employment. The exchange rate, following the impact effect, will appreciate, or depreciate, depending on (A12).

References

Black, S. *International Money Markets and Flexible Exchange Rates.* Princeton Studies in International Finance, no. 32. Princeton, N.J.: Princeton Univ. Press, 1973.

———. "Exchange Rate Policies for Less Developed Countries in a World of Floating Rates." Mimeographed. Vanderbilt Univ., 1975.

Dornbusch, R. "Exchange Rate Expectations and Monetary Policy." *J. Internat. Econ.* (1976), forthcoming (*a*).

———. "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy." *Scandinavian J. Econ.* 2 (May 1976): 255–75 (*b*).

Fischer, S. "Stability and Exchange Rate Systems in a Monetarist Model of the Balance of Payments." In *The Political Economy of Monetary Reform*, edited by R. Aliber. London: Macmillan, 1976.

Fleming, M. "Domestic Financial Policies under Fixed and Floating Exchange Rates." *I.M.F. Staff Papers* 9 (November 1962): 369–79.

Frenkel, J. A. "A Monetary Approach to the Exchange Rate." *Scandinavian J. Econ.* 2 (May 1976): 200–221.

Henderson, D. "Monetary, Fiscal and Exchange Rate Policy in a Two-Country, Short-Run Macroeconomic Model." Mimeographed. Board of Governors, Federal Res., 1975.

Kouri, P. *Essays on the Theory of Flexible Exchange Rates.* Ph.D. dissertation, Massachusetts Inst. Tech., 1975.

———. "The Exchange Rate and the Balance of Payments in the Short Run and in the Long Run." *Scandinavian J. Econ.* 2 (May 1976): 280–304.

Mundell, R. A. "Exchange Rate Margins and Economic Policy." In *Money in the International Order*, edited by C. Murphy. Dallas: Southern Methodist Univ. Press, 1964.

———. *International Economics*. New York: Macmillan, 1968.

Mussa, M. "The Exchange Rate, the Balance of Payments and Monetary and Fiscal Policy under a Regime of Controlled Floating." *Scandinavian J. Econ.* 2 (May 1976): 229–48.

Niehans, J. "Some Doubts about the Efficacy of Monetary Policy under Flexible Exchange Rates." *J. Internat. Econ.* 5 (August 1975) 275–81.

An Alternative Approach to the Analysis of Taxation

Yoram Barzel

University of Washington

Because commodities as transacted are complex, tax statutes could not cover all margins subject to optimization. A tax will induce, then, substitution within the commodity away from the taxed attributes and into the others. The results of a test on cigarettes are consistent with our prediction that the effects of unit and ad valorem taxes will differ both from each other and from those predicted by the conventional model. It is shown that, although the market will adjust in numerous changeable characteristics, the adjustment is constrained by the condition that the sum of the dollar value of the inefficiencies and of the tax paid is minimized.

> The Dog and the Accountant
>
> A dog whose fleas had become unbearable sought the advice of an accountant.
>
> "Can I claim these fleas as exemptions?" he asked.
>
> The accountant smiled and shook his head. "Under certain conditions," he explained, "ticks can be claimed as blood relatives. Fleas, on the other hand, are considered luxuries and, far from being deductible, are actually subject to tax."
>
> The following day, the dog had no more fleas; or he didn't declare them, which comes to the same thing.
>
> Moral: There can be no taxation without misrepresentation.
>
> [J. B. Handelsman][1]

Thanks are due to numerous individuals for their comments, particularly to Dan Benjamin, Steve Cheung, Bill Schworm, and Dan Wisecarver. The paper was written, in part, during a visit at the Hoover Institution.

[1] Reprinted by permission, © 1973 by The New Yorker Magazine, Inc.

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved.

This paper offers an extension of the traditional approach to the analysis of taxation capable of explaining observations attributable to taxation which were not previously explicable. Underlying the change in approach are the premises that *every* commodity is composed of a set of characteristics whose levels are economically determined and that no tax will be so comprehensively specified as to control all the characteristics of a commodity Market adjustments to a tax will take place not only in the quantity of the good but also in the make-up of its characteristics

By standard analysis, a tax on the tonnage of ships would simply lead to a substitution resulting in lower tonnage Such analysis, however, is incapable of explaining why "shipowners had vessels built excessively long, deep and narrow, 'all more or less crank' [i e , poor in seaworthiness]" (Lane 1964). In the above instance, the tax definition of tonnage (motivated, it seems by considerations of the costs of measurement) was basically proportionate to the length of ships and negatively related to the square of their width. Although deeper and narrower ships took a toll in handling, they did achieve from 1 1 to 1 4 tons of carrying capacity for each "ton" as measured for the purpose of tax

In this paper it will be shown that, subject to the relevant constraints, the losses to market participants resulting from either unit or ad valorem taxes are treated as a magnitude to be minimized, rendering the adjustments in characteristics predictable

I. The Setting of the Problem

A consumer's evaluation of a commodity rests on its various characteristics The commodity "gasoline," for example, is demanded for its potential for generating energy, for its octane rating, and for its additives, as well as for such peripheral services as windshield cleaning The choice open to consumers of commodities composed of many characteristics has been analyzed by Griliches (1961) in his empirical work on automobile prices and by Lancaster (1966) in his theoretical work on demand We wish to proceed further by attempting to explain how the market will determine the levels of a commodity's characteristics in response to a tax

For any given commodity tax, it is crucial to know how the statute defines the unit of the commodity, how that unit is measured, and how the tax is to be enforced. The unit for tax purposes may correspond to the market norm—gallons of gasoline, or packages of 20 cigarettes Granted that the legal definition of the unit will certainly be more detailed, it is still highly unlikely to cover all revelant characteristics[2] Furthermore, entire characteristics are subject to inclusion or exclusion with varying economic conditions. We seek to discover the nature and extent of these adjustments. As will become evident, the results depend critically on whether the tax is fixed per unit of the commodity or is ad valorem.

---

[2] The legal definition will almost inevitably become more detailed as inadequacies become evident For example, the present tax-purpose definition of a cigarette will presumably expand eventually to cover the newly emergent "little cigars." Earlier, a maximum length for cigarettes had to be specified to counter the marketing of extra-long cigarettes, taxed at a constant rate per unit, which could subsequently be cut into standard lengths.

With a few explicit exceptions, we assume that prior to the imposition of a tax the market is competitive and free of distortions. A partial equilibrium framework is used throughout the analysis, and we abstract from possible feedbacks due to the effect of the tax on the level of income or on income distribution.

Because existing studies of the effects of commodity taxes employ demand and supply elasticities estimated prior to the tax, their results reflect only the change in volume of the commodity as a whole with no allowance for change in the mix of characteristics or any resultant effect on allocation. In a more careful analysis, this bias would be avoided by the treatment of different varieties as separate commodities, taking into account not only the individual elasticity of each but also their cross-elasticities, as noted by Harberger (1964, pp. 25–70). While this is a step in the right direction, it still does not allow for possible tax-induced changes in the characteristics of each variety. ("Cigarettes," for instance, may shift through a wide spectrum: the quality of tobacco and of its wrapping, the per-package retail services afforded for a given brand, the menthol or other components, and, of course, the length of the cigarettes themselves.) This paper explicitly takes into account this capacity for internal change within each commodity.

II. The Constant per Unit Tax

The traditional theory of taxation under competition is based on the comparison between the systems $D(P) - S(P) = 0$, and $D(P^T) - S(P^T - T) = 0$, where $D(P)$ is the market demand for a commodity, $S(P)$ is its supply, P is the price per unit before the tax, P^T after the tax, and T is the tax per unit. Misallocation occurs because resources shift away from the taxed industry into less highly valued alternatives. For the marginal unit the value of the misallocation is equal to the tax rate. Under common but rather restrictive assumptions, the economic waste as a result of the tax is measured by $\frac{1}{2}[\eta\varepsilon/(\eta + \varepsilon)](T/P)^2$ where η and ε are the respective demand and supply elasticities.

For a commodity composed of several characteristics, equality between the marginal valuation of the commodity as a whole and its marginal cost is a necessary, but not a sufficient, condition for competitive equilibrium. The marginal valuation (MV) of each of the characteristics must equal its marginal cost (MC).[3]

---

[3] In what follows, we assume that each of the marginal valuations is falling and each of the marginal costs is rising.

Consider a commodity possessing n characteristics. At the pretax equilibrium $MC_i = MV_i$, $i = 1, \ldots, n$. Defining

$$MC = \sum_{i=1}^{n} MC_i \quad \text{and} \quad MV = \sum_{i=1}^{n} MV_i$$

for the commodity as a whole,[4] the equality for each characteristic implies also that $MC = MV$.[5] After a tax-induced reduction in number of units, but prior to any change in the levels of the characteristics per unit, the following equality will hold

$$\sum MC_i^T + T = P^T = \sum MV_i^T. \tag{1}$$

The reduction in quantity of the commodity as a whole constitutes a downward move along each of the MC_i schedules and an upward move along each of the MV_i schedules, so that

$$MC_i^T < MC_i = MV_i < MV_i^T \qquad \text{for all } i\,^6 \tag{2}$$

Suppose now that the tax statute defines the commodity in terms of the pretax equilibrium levels of characteristics, $i = 1, \ldots, k$, but that characteristics $k + 1, \ldots, n$ are not specified in the tax statute. We can decompose (1) into

$$\sum_{i=1}^{k} MC_i^T + \sum_{i=k+1}^{n} MC_i^T + T = P^T = \sum_{i=1}^{k} MV_i^T + \sum_{i=k+1}^{n} MV_i^T. \tag{1'}$$

Given (1′) and the inequalities in (2), prior to any change in the characteristics per unit of the commodity we have

$$\sum_{i=1}^{k} MC_i^T + T > \sum_{i=1}^{k} MV_i^T, \tag{3a}$$

and

$$\sum_{i=k+1}^{n} MC_i^T < \sum_{i=k+1}^{n} MV_i^T. \tag{3b}$$

Here (3a) relates to characteristics specified by the tax statute and (3b) to those which escape the constraint. The tax, strictly speaking, is imposed only on characteristics $1, \ldots, k$. Since the tax-inclusive cost of the first k characteristics exceeds their valuation, while the converse holds for the other $n - k$ group, there is incentive to contract the former relative to the latter. Since the tax statute determines the level of the former per unit of the good, adjustment will have to take place by reducing the number of units produced while increasing the levels of the uncontrolled characteristics per unit of the good. Expansion of the untaxed characteristics constitutes an effective substitution against the tax. Adjustment will continue to the point where the inequality in (3a) and all of the $n - k$ components in the inequality in (3b) turn into equalities.

---

[4] To fix units of measurement we set the pretax level of each of the characteristics at one unit per unit of the commodity. Thanks are due John Floyd for correction of a previous error.

[5] An MV_i should be construed as the change in the total value of the commodity induced by a unit change in the ith characteristic and similarly for MC_i.

[6] We abstract from cross-effects among the characteristics in either cost or valuation functions which could conceivably reverse the inequalities.

Since characteristics $1, \ldots, k$ are constrained to a maximum level by the tax statute and since there is basically no incentive to raise them, their level will stay unchanged and we can consider them as one. Characteristics $k + 1, \ldots, n$ are not constrained, but to simplify we shall assume that their relative levels are maintained. Let us call the taxed set of characteristics A, the untaxed B, and the commodity as a whole C.

The tax will cause a decline in the number of units of A demanded. The demand for B may increase or fall, depending on whether A and B are substitutes or complements, but that of B relative to A will have to increase.[7] While each unit of C will contain the same A as before, the tax will cause it to include more of B. In other words, the tax will induce an upgrading in the "quality" of C inasmuch as each unit will contain more B.[8]

The opportunity to modify the composition of the commodity reduces the tax burden implied in the fixed-quality model, and the equilibrium quantity will increase correspondingly. The "welfare triangle," therefore, is smaller than in the fixed-quality case. The change in resource allocation induced by the tax, however, is no longer confined to a reduction in quantity of the taxed good but now includes the costs of producing an inefficient mix of characteristics within the good.[9]

Thus participants in the market for the commodity suffer three kinds of losses from the tax: the tax payments,[10] the welfare triangle from reduced production, and the loss of efficiency in production. Consumers and producers of the taxed commodity will benefit most by minimizing the sum of these losses. Our central proposition is that market forces will actually produce that result

---

[7] In a two-commodity world and holding individual incomes constant the demand for B per unit of A will remain unchanged only if they are perfect complements. Even though basically the tax will not affect aggregate income, it will affect its distribution. The individual income effects will not necessarily cancel out and may produce a perverse outcome. In addition, as noted by Gould and Segall (1969), interaction with other commodities could conceivably upset the results. Finally, the change in the quantities of A and B implies movements along the (pretax) supply schedules. The two supply elasticities may differ sufficiently to change the expense of A relative to B. None of these possibilities seems empirically important. In what follows, we abstract from all of them.

[8] The term "quality" is not used in a normative sense. Indeed, if B rather than A were being used as a unit of measure the "quality" would be reduced by the lesser proportion of A. It is used here as a shorthand notation for the quantity of given characteristics per unit of the commodity.

[9] A more complete discussion of these points in relation to a special case is presented in Section V, following.

[10] As already mentioned, we abstract from whatever benefits may arise from the use to which the tax is put.

To demonstrate the proposition consider the market for our commodity in a (taxless) economy in full competitive equilibrium From all possible positions open to the economy, equilibrium is at the point at which the sum of consumers' and producers' surplus is maximized. Although most economic models consider one-dimensional goods, the fundamental properties of equilibrium are not affected when commodities have more than one characteristic.

With any change in the price of a factor used in producing the commodity, the economy will move to a new equilibrium Suppose now a change occurs which has no effect on *B* but which raises by *T* the real cost of producing *A*. From the previous analysis of the effect of the tax (and abstracting from resultant differential changes in the rest of the economy) it is clear that the same equilibrium will be reached at the new factor price as would have occurred under the tax. Given the change in cost, the new equilibrium is again a position of maximum surplus, representing the smallest possible loss as compared to the prior situation. The market will react to the higher factor cost by economizing in its use (1) by contracting the output of the commodity itself and (2) by substituting toward those of its characteristics which use relatively less of the high-priced factor In short, the market reaction is identical to that in the tax case, in both instances the adjustments constitute a minimization of the sum of the two cost components. Thus the proposition is demonstrated that following the imposition of a tax the market will move to minimize the sum of the tax paid and the welfare loss

This proposition has been set out in terms of only two characteristics, but its logic is equally applicable irrespective of their number At competitive equilibrium for any such commodity, economic surplus is maximized. Suppose an increase in the price of a factor used to produce some, but not all, of the characteristics takes place. The levels of each will be so adjusted that all the marginal conditions will be satisfied The new equilibrium, subject to the higher factor price, will again be a position of maximum surplus, and the move to the second equilibrium will have entailed the smallest possible loss in surplus. Given the complete analogy of impact between a tax and an increase in factor price, the fundamental result enunciated above is entirely general.[11]

---

[11] Such terms as welfare triangle, misallocation, inefficiency, distortions, and the like carry normative connotations We have just shown, however, that they possess positive properties as the subject of market minimization This terminology, ingrained in the literature, is retained here but is used only in its positive sense.

III. The ad Valorem Tax

Suppose an ad valorem tax of rate t is imposed on a commodity having n characteristics The tax statute here defines the commodity in terms of the *minimum* levels of characteristics 1, , k The after-tax equilibrium prior to any change in characteristics can be written as

$$(1 + t) \sum \mathrm{MC}_i^t = P^t = \sum \mathrm{MV}_i^t, \tag{4}$$

which is the counterpart of (1) for the constant-unit tax Expression (4) can be decomposed, as is (1′), into the constrained and unconstrained components

$$(1 + t) \sum_{i=1}^{k} \mathrm{MC}_i^t + (1 + t) \sum_{i=k+1}^{n} \mathrm{MC}_i^t = P^t = \sum_{i=1}^{k} \mathrm{MV}_i^t + \sum_{i=k+1}^{n} \mathrm{MV}_i^t \tag{4'}$$

Economic choice again determines both what characteristics will constitute a given commodity and their levels A characteristic will be included in a commodity only if it is cheaper when part of the package than when sold separately Since characteristics $k + 1$, , n are not constrained, reduction of their levels will serve to evade some of the tax. Most important, these characteristics are taxed only so long as they are sold as part of the commodity[12] Thus, if the pretax advantage of selling the $k + 1$st characteristic as part of the package exceeded that of selling it separately by less than $tx\mathrm{MC}_{k+1}$, it may be removed altogether from the package when the tax is imposed

To simplify, let us again lump the first k characteristics into A and the remainder into B, and assume that A and B are consumed in fixed proportions. Suppose that, prior to the tax, a package was preferred This is shown in figure 1, where D is the demand for the package, $S(P)$ is the supply of the package, $S(A)$ is the supply of one component, and $S(A + B)$ is the sum of the supplies of the two components (plus assembly, if necessary)

What would be the effect of an ad valorem tax? If the definition of the commodity for tax purposes is $A + B$, then the conventional analysis

---

[12] In at least some situations, substantial savings accrue to buying a package rather than its individual components A study published in *Money Magazine* (1972) shows that while the retail price of a Ford Maverick was then \$2,506.55 the retail price for its more than 10,000 parts, unpainted and unassembled, would have totaled \$6,435.22. To this could be added the cost to the buyer of time spent in searching and buying.

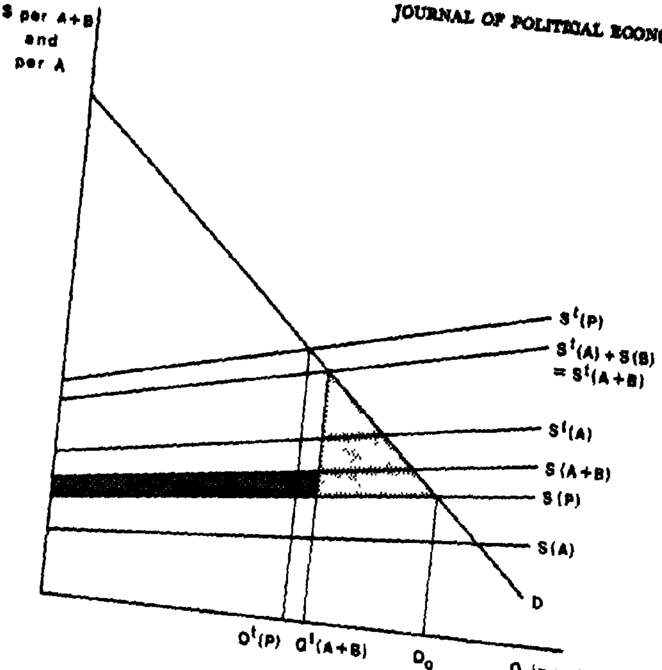


FIG 1

holds. But if the definition covers A only,[13] the cost of the package rises relative to that of its components, since B is taxed only when sold as part of the package. $S^t(A + B)$ is the sum of the supplies of the components of which only A is taxed and $S^t(P)$ is the supply of the taxed package. As drawn, the former intersects the demand curve at a lower point and so is preferred. The price for $A + B$ paid by the consumer is less if he buys the components separately, so the after-tax quantity falls only to $Q^t(A + B)$ rather than to $Q^t(P)$, which would be the equilibrium quantity were the whole package taxed. As a result, the welfare triangle is also smaller than when quality is unchanged But the reduced output is now produced at a resource cost according to $S(A + B)$ rather than to the lower $S(P)$, and this inefficiency adds to the welfare cost of the tax. The two components of the total welfare loss form the shaded areas in figure 1.

Tax revenue is now obtained from a larger number of units—$Q^t(A + B)$ rather than $Q^t(P)$; but as the tax base is smaller, the tax per unit is less and we cannot determine a priori the direction in which the total tax payment will change. The after-tax price of the commodity as defined for tax purposes not only will be less than predicted by the constant-quality model but may actually fall below the pretax price[14]

---

[13] Such would be the case where a new air-conditioned car is taxed at the same rate as one lacking that option and where separately purchased air-conditioning units are untaxed. For the tax to apply to all components of a car, the statute would have to cover each of many thousand parts, specifying for each a minimum standard of quality.

The analysis does not change in any essential way when we drop the assumptions of only two attributes consumed in fixed proportion. An ad valorem tax will induce a reduction in all the unconstrained attributes; some will be retained as part of the commodity but at lower levels, while others will be dropped altogether.[15]

IV. The Distinction between Unit and ad Valorem Taxes

Provided both that the appropriate tax rates are imposed and that quality is held constant, when a commodity is taxed the subsequent resource allocation and tax revenue will be the same regardless whether the tax is constant per unit or ad valorem But when quality is not fully controlled, the effects of the two types of taxes differ radically. The excise tax induces a substitution of quality for quantity, resulting in a greater increase in the price per unit[16] than predicted by the constant-quality model. The ad valorem tax induces a reduction in quality and a lower after-tax price than predicted Thus the statement by Due (1959, p. 298, italics added) that "this difference [between ad valorem and specific taxes] does not alter the *basic* initial reactions or the *ultimate* distribution of burden" has to be rejected if our hypothesis is valid

It is evident that when quality is adjusted, neither the per unit nor the ad valorem tax remains truly constant. The former varies because the taxable unit has been changed. The latter is transformed into a variable tax because the value of the product is altered. In each case, the motivation for the quality change is a lower tax per unit of service.

The relation between the two taxes can be traced to an essential feature of commodities transacted in the market Consider again a commodity possessing n characteristics and assume that the tax statute defines the product in terms of the first k. If deletion of characteristics were costless, the reaction to either a per-unit or an ad valorem tax would be the same. Characteristics $k + 1, \ldots, n$ are untaxed by the excise tax; their removal from the commodity will exempt them from the ad valorem tax; thus substitution toward these characteristics is implied in either case We can predict a reduction in the consumption of the first k characteristics, and a relative increase in the consumption of the others.

---

[14] The corporate income tax is fundamentally an ad valorem tax. When such a tax is imposed or raised, it can be evaded in part by switching to corporate financing based more on debt and less on equity The risk borne by equity holders will then increase. The measured rate of return to equity, reflecting this risk, might exceed the change in the tax rate, giving the impression of more than 100 percent shifting.

[15] It may appear that a general sales tax on final consumer commodities (or a uniform income tax) will avoid these problems However, in addition to the difficulty of taxing home production, there is an incentive for employers to provide their employees with consumption goods such as air conditioning and uniforms—classified under the tax as intermediate and not subject to tax—thereby still partly evading the tax.

[16] The unit referred to is that defined for tax purposes.

Suppose now that the ad valorem tax is set at such a rate that the amount realized after the elimination of characteristics will equal that attainable under a unit tax. The two taxes would now have *identical* effects on resource allocation. If we measure the commodity in units as defined by the tax statute, the tax-inclusive price will be lower for the ad valorem tax, but this is simply because the characteristics $k + 1, \quad , n$ are sold separately

Of course, it is invalid to assume that the deletion of characteristics will be costless, since they would not originally have been incorporated in the commodity unless this offered advantages. An ad valorem tax will still lead to a reduction of the levels of all the $k + 1, \quad , n$ characteristics per unit of the good While some characteristics may be completely eliminated, others will merely be reduced The two taxes, then, will result in quite dissimilar effects on resource allocation as well as in different observed prices and tax revenues [17]

V. Tax and Durability

In this and the following section, the analysis will be confined to a commodity with one continuously variable attribute—durability This simplification not only will clarify some of the foregoing conclusions, but will help in the analysis of additional relevant issues Thus, for the commodity selected for illustration—light bulbs—we consider only the variable of durability The bulbs are assumed to be homogeneous in such dimensions as strength and color, and while they last they provide a constant-quality flow of services (the one-horse-shay model)

Consumers, we assume, value the flow of light services but not durability per se. A consumer, then, would be willing to pay the same price for 12 1-month bulbs as for a single 1-year bulb. This valuation scheme becomes plausible if we conceive of the consumer as renting the bulbs, that is, purchasing the light service rather than the product The cost of changing bulbs is then transferred to the seller and can be incorporated in his cost function.[18] Under these conditions the durability offered in the market will be that for which the price of light services is lowest, satisfying the condition that the cost of an infinitesimal increase in volume of service will be the same whether obtained by an infinitesimal increase in the number of bulbs or in their durability.

---

[17] It would be interesting to discover under what conditions one type of tax would be selected over the other.

[18] We also assume that for any given volume of service there is always a single durability for which the cost is lower than for any combination of durabilities.

Now consider an excise tax on light bulbs We assume that the tax statute does not constrain durability This is not unreasonable, since in our model durability is not an explicit feature of bulbs as only one durability is provided by the market at a point in time A tax which is constant per bulb raises the marginal cost of bulbs by the amount of the tax but does not affect the marginal cost of durability The least-cost method of obtaining any given output of light services after the tax is to increase durability to the point where equality on the two margins is restored

When we compare the constant durability model—fundamentally the accepted model—with the unconstrained durability model, several major differences emerge With the removal of the constraint, the tax-inclusive cost of service will be lowered and the quantity of services increased In general, competition drives the price to the lowest level consistent with cost. Here a move to a higher durability reduces the tax-inclusive cost per unit of service and competition will guarantee that more durable bulbs will actually be provided

Received theory predicts that an excise tax will result in an increase in price not larger than the rate of tax However, since the increase in durability will itself tend to increase the price of light bulbs, the actual increase in price per light bulb may easily exceed the rate of tax.[19]

Much of the content of the above results can be presented diagrammatically In figure 2, the units on the horizontal axis represent light services and on the vertical axis, dollars per unit of service Since the demand for service is independent of durability, a single demand curve suffices For each durability, however, there is a different cost curve To simplify graphic exposition, it is now assumed that cost curves for varying durabilities do not intersect We denote by $S(x)$ the cost curve for durability x Two such cost curves are drawn in figure 2 Suppose that the lower one, $S(2)$, provides light services at lower resource cost than any other durability, either shorter or longer Before taxation, then, $Q(2)$ and $P(2)$ are the market quantity and price

A tax of T per light bulb is now imposed Each of the $S(x)$ curves will shift upward by amounts inversely proportional to x, the tax per unit of service on a 4-month bulb is half that on a 2-month bulb The lowest curve will now be one of longer durability, the higher the tax, the greater

---

[19] Following Harberger (1964), one could alternatively argue that the tax is actually imposed not on a single commodity but rather on a whole array of commodities where each durability constitutes a separate commodity It is clear that the tax-induced increase in the consumer price for each would not exceed the rate of the tax. But the elasticity of substitution in consumption between alternative durabilities is such that the old durability is entirely replaced. This is formally correct but does not seem useful. In most empirical studies of taxation, the level of aggregation of the taxed commodity is quite high, indicating lack of awareness of the problem of tax-induced substitution within what is classified as a homogeneous commodity.

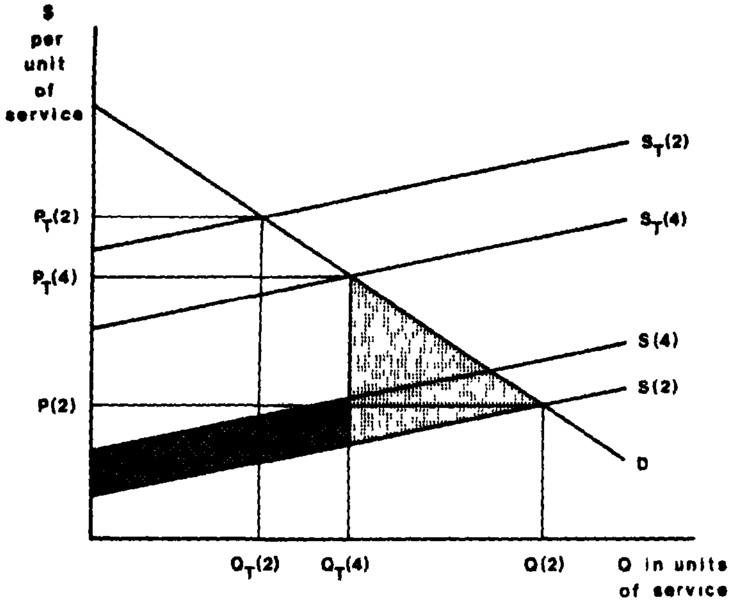


FIG 2

that durability. We denote by $S_T(x)$ the tax-inclusive cost of light services obtained from bulbs of durability x, and suppose that among all these $S_T(4)$ is lowest.

After the tax, $Q_T(2)$ and $P_T(2)$ are the equilibrium quantity and price if durability is constrained to 2 But if durability is not constrained, the final price per unit of service will be lower and the final quantity of services higher, since $S_T(4)$ lies below $S_T(2)$ Note, however, that the price observed in the market is not per unit of service but per bulb Thus even though $P_T(4) < P_T(2)$ and $P_T(2) - P(2) <$ tax, the increase in the price per bulb may exceed the tax rate because the after-tax bulb is more durable. Indeed, if the supply curves are perfectly elastic, the increase in the price per bulb will *necessarily* exceed the rate of tax.

Let us now turn to tax receipts. In figure 2 these are represented by the product of the after-tax market quantity and the difference between the after-tax market price and the pretax cost. It is easy to see that tax receipts may go up or down when we compare the constrained with the unconstrained durabilities. For instance, if the tax is originally set to yield maximum tax revenue while constraining durability to its pretax level, when the constraint is relaxed the substitution towards higher durability will necessarily result in a revenue loss.[20] On the other hand, if the imposed tax rate is just prohibitive, it will not remain so for certain longer durabilities, and tax revenues will then increase from zero to some positive value.

---

[20] Finding the tax rate that yields maximum revenue is equivalent to finding the largest rectangle that can be inscribed between the demand and supply curves to the left of their intersection. Since higher durabilities reduce the area between the two curves, the new largest rectangle has to be smaller than the old one.

What are the pertinent welfare costs for the two alternatives? With no change in durability, the deadweight loss due to the tax is the conventional triangular area between the demand curve and the supply curve $S(2)$ from $Q_T(2)$ to $Q(2)$. When durability is permitted to adjust, the price paid by consumers is lower and the quantity purchased larger than when durability is constrained Thus the conventional welfare triangle is reduced to the shaded area between D and $S(2)$ from $Q_T(4)$ to $Q(2)$

However, when durability is variable, misallocation is no longer restricted to that resulting from contraction of output Two-month durability is most efficient since it provides light services at lowest resource cost, but in the process of partially evading the tax durability has now been increased to 4 months, thus the *entire* quantity is being produced inefficiently The quantity $Q_T(4)$ could have been produced at the resource cost indicated by $S(2)$ rather than by $S(4)$ The shaded area between $S(4)$ and $S(2)$ up to quantity $Q_T(4)$ represents the value of the tax-induced excess resources used in producing light services

What are the properties of the new equilibrium position? Together, producers and consumers search for that durability for which the effect of the tax is minimized that is, for the real price that results in the smallest contraction of service output In other words, market participants search for the durability that leads to the lowest after-tax price per unit of service The tax paid per light bulb is still T, but now a bulb provides more services, thus the real tax is lower by a corresponding amount.

To depict further the properties of the after-tax equilibrium, figure 3 adds to the demand and supply curves of figure 2 the supply curves of 3-month bulbs, both before and after the tax and a new curve denoted H.[21] Each point on H represents a net-of-tax market equilibrium position if durability is constrained to a particular value Thus, point 2 is the after-tax equilibrium to suppliers if durability is constrained to 2 months. Similarly for points 3 and 4 Points 2′, 3′, and 4′ are the corresponding points of demand That is, given the tax and the constrained durabilities, these are the optimal points to consumers. The distance between 2 and 2′, for instance, is the tax per unit of service for a bulb of 2-month durability. The H curve is tangent to $S(2)$ at point 2, since $S(2)$ is the lowest supply curve.

As we move from point 2 through higher durabilities, output is produced at increasing resource cost. In the durability range from 2 to 4 months, however, this increase in cost is more than matched by tax savings. In that range, then, the price paid by consumers is declining, the quantity purchased rises, and the welfare-loss triangle shrinks correspondingly.

---

[21] The discussion of figure 3, and a following related consideration of subsidies, were stimulated by a conversation with Al Harberger.

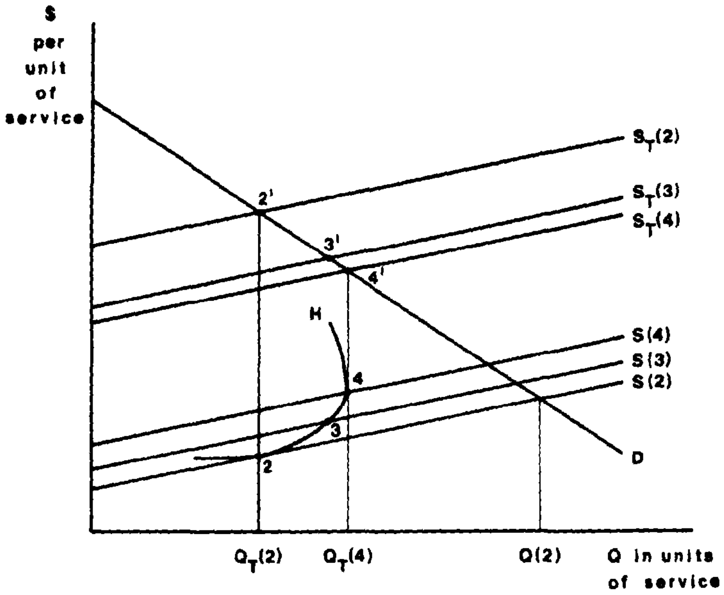


FIG. 3

Beyond point 4 the tax saving of increased durability is less than the extra resource cost At point 4, where durability is identical to that chosen in the absence of constraint, the *H* curve is vertical and the incremental gain from a further saving in taxes is exactly matched by an increase in cost of production. That point, then, satisfies the first order condition for a minimum Since point 4 is market determined, the earlier proposition that the market will minimize the dollar value of the sum of welfare losses and tax paid is again demonstrated, though from a different angle.[22]

VI. Subsidies

It may appear that a discussion of subsidies at this point is superfluous Aren't subsidies the "mirror image" of taxes? Why would not a subsidy simply reverse the signs of the tax case? It is true that a tax is a negative subsidy, the effect of the former is to reduce market quantity and of the latter to increase it. But the effect is not symmetric with respect to resource allocation. *Both* lead to positive welfare triangles and to inefficient moves from the optimal mix of characteristics into mixes which use up more

---

[22] In the case just discussed, would an ad valorem tax affect durability, or would it be "neutral"? If the tax is on bulbs, but not on their installation, the buyer has a stronger incentive to install his own bulbs. As the cost of installation relative to bulbs is then lower, durability will be shortened.

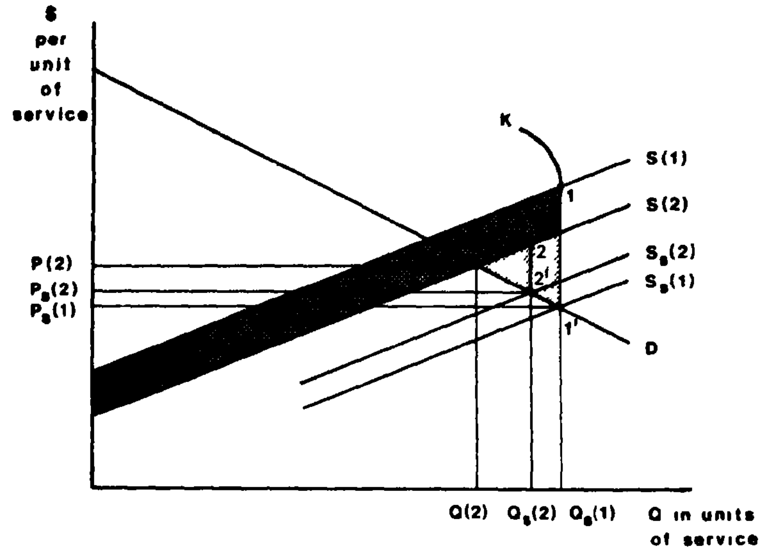


FIG 4

resources In the remainder of this section we examine this aspect of the effects of a constant per unit subsidy on a durable commodity.

In addition to directly increasing quantity, a subsidy will lead to an added expansion in quantity because of substitution against durability, since the price per unit of service is thus reduced. In figure 4, $S(2)$ and D are again the same as in figure 2 A subsidy, S, is granted per light bulb of whatever durability If durability is constrained to remain at 2 months, the price paid by consumers per unit of service will decline to $P_S(2)$ and the quantity will increase to $Q_S(2)$ But for durabilities shorter than 2 months the subsidy per unit of service is higher, thus light service will be cheaper Suppose that among bulbs of shorter life those of 1-month durability are best, $S(1)$ and $S_S(1)$ as drawn in figure 4 show that when the durability constraint is removed the price per unit of service will fall to $P_S(1)$ while the quantity of service consumed will increase to $Q_S(1)$

In regard to welfare loss, since the market quantity of services is lower when durability is held constant, the removal of that constraint leads to a *larger* welfare triangle—the shaded triangle in figure 4. In addition, bulbs of 1-month durability provide light service at a higher resource cost than 2-month bulbs, consequently the entire output is produced inefficiently The total resource cost of this inefficiency is the area between $S(1)$ and $S(2)$ to quantity $Q_S(1)$, which is also shaded in the diagram.

It is evident that if the rate of the subsidy just discussed equals the rate of the tax considered earlier, and given linear demand and cost curves, the adjustment to the subsidy is larger than that to the tax.[23] If the tax and subsidy rates are the same, the changes which will occur in price and quantity under conditions of constrained durability are of the same size (in absolute terms). When durability is allowed to change, a downward move is induced along the demand curve by either the tax or the subsidy. This mitigates the initial effect of the tax but enhances that of the subsidy— which thus has the greater total impact.

More generally, a series of constant tax doses will result in successively smaller adjustments in the price and quantity of service. The intuitive explanation for this is that since each tax increment leads to the production of more durable bulbs for the market, the new levy has a smaller proportionate effect.[24]

The curve K in figure 4 is obtained similarly to H in figure 3. This curve, too, is tangent to S_2 and is vertical at output level $Q_S(1)$ If durability is constrained to 2 months, producers will operate at point 2 after the subsidy is granted As they are permitted to produce shorter durabilities they will move along the curve K as far as, but not beyond, point 1. At that point the benefit from obtaining a larger subsidy by switching to a still shorter durability is exactly nullified by the increase in cost of the shorter durability Point 1 produces the largest welfare triangle; as durability was reduced from 2 months to 1, the triangle was increasing continuously, but it then begins to shrink because still shorter durabilities lead to a decline in market quantity. At point 1, which will be chosen by the market in the absence of constraint, the sum of welfare loss and the (negative) tax paid is at a minimum—a small move leaves the area of the triangle unchanged while the change in welfare loss due to inefficient production is exactly matched by the change in subsidy.

Consider now a change away from point 2 Here, an infinitesimal reduction in durability will result in an increased subsidy, but the efficiency in production will not be affected. This, then, will lead to a fall in price paid by consumers, to a corresponding increase in output, and hence to a larger welfare triangle. It is evident, then, as stated earlier, that any move to a shorter durability must result in an unambiguous increase in welfare cost.

When it is possible to tighten the legal definition of the commodity with the objective either of increasing tax revenue or of reducing the dollar value of the misallocation, more information is needed to determine which way the tightening will work in the case of a tax. With a subsidy, on the other hand, a move to prevent substitution among qualities will always reduce both subsidy payments and misallocation.[25]

---

[23] Of course, a simultaneous tax and subsidy of the same size will identically cancel out. Here the starting point of the subsidy is the end point of the tax. In the textual comparison, the starting point for both is the same.

[24] Note, however, that the observations in empirical work typically correspond to those on light bulbs rather than on light services It is not entirely clear, then, that the observed price increases will continue declining with successive equal tax doses; a result which anyhow depended on the assumption of linearity.

VII. A Test of the Hypothesis

Our fundamental hypothesis is that the effect of a tax on the price of a commodity, as measured, will differ from that indicated by conventional analysis. Specifically, we predict that a tax which is constant per unit will increase consumer price, as normally measured, more than is predicted by the received model, whereas an ad valorem tax will raise price less than a comparable per-unit tax.

Consider also the "mark-up" hypothesis which states (inconsistently with demand theory) that when the price to the seller is increased by whatever reason, he will raise the retail price by a larger amount. We argue that the effect produced by an ad valorem tax will differ radically from that of a constant per unit tax. Data on the effects of the two taxes may enable us to test the received model and also to choose between the mark-up hypothesis and ours

Reported in the following paragraphs are tests of the hypotheses using data on cigarettes [26] Starting in 1954, the Tobacco Tax Council (1972) has compiled and reported the retail price of cigarettes on a state-by-state and year-by-year basis [27] The "prices" are actually within-state indices using a set of fixed weights for all states in all years Weights are given to three types of cigarettes—regular, king, and filter (both 85 mm and 100 mm)—and to three types of sale—single pack over the counter, vending machine, and carton lot [28]

Cigarettes provide an almost ideal setting for testing our hypotheses. The bulk of cigarettes originate from around North Carolina and are distributed nationally. Apart from the effect of shipping costs (which is possible to control for), the observed price differences at a point in time are presumably due largely to state taxes

While a federal tax remained constant at 8 cents per package for the entire period, state taxes varied greatly within states and over time. In 1972, the most recent year used, these ranged from 2 cents per package in North Carolina to 21 cents in Connecticut. In the same states the taxes had been, respectively, 0 and 3 cents in 1954.

---

[25] Incidental to the central theme of this paper is the hypothesis that the quality constraints actually imposed would be more severe for a subsidy than for a tax.

[26] Attempts to test our hypothesis against data on gasoline and liquor are reported in the Appendix.

[27] For most years, the data are given as of November 1. For 1957, 1958, and 1959, the date used is October 1.

[28] Information provided in a letter from the Tobacco Tax Council.

TABLE 1

CIGARETTE REGRESSION RESULTS*

Variable	Coefficient	Standard Error	*t*-Value
Constant	14.801	0.201	73.564
Tax (in cents per package) .	1.065	0.011	101.439
New Hampshire dummy	−0.072	0.014	−5.078
Distance (in 100 miles)	0.011	0.004	3.072
	$\bar{R}^2 = .93$		

* A set of year dummies included in the regression showed an upward drift from −1.9 (cents per package) in 1954 to 0.8 in 1970 declining back to zero (the "base" year) in 1972.

It seems reasonable to expect that the supply elasticity of cigarettes to a single state is very high. Thus if the conventional model is correct, a 1-cent difference in tax per package between two states will result, *ceteris paribus*, in close to a 1-cent difference in price. Using our model, we predict a price increase larger than 1 cent if the tax is constant per unit and a smaller increase if it is ad valorem.

Table 1 presents results from a regression based on 931 observations.[29] The dependent variable is the retail price, deflated by the consumer price index for October of the given year. The coefficient of the tax (also deflated by the consumer price index) indicates that a 1-cent increase in tax raises retail price by 1.065 cents, which is significantly larger than one, as the *t*-ratio for that test exceeds 6.[30] The received notion that the price will increase at most by the amount of the tax can be rejected.[31]

In New Hampshire the tax is ad valorem. The regression coefficient for the dummy variable for that state, measuring the differential from other states with the comparable unit tax, is −0.072 and its standard error is 0.014. The retail price in New Hampshire, then, is significantly less than that predicted for a state with a comparable per package tax,[32] and the hypothesis can be rejected that comparable per unit and ad valorem taxes will have identical effects.[33] Also rejected is the mark-up hypothesis, as it fails to predict differential effects from the two types of tax Finally, the coefficient of the distance variable, inserted to control for the differential shipping costs from North Carolina, is positive, as expected, and significant

---

[29] These are one observation each for 19 years (1954–72) for the District of Columbia and for each of the 48 continental states. Alaska and Hawaii, for which data are available for only part of the period, were deleted from the sample because of the ambiguity of the distance measure. Note that Hawaii imposes an ad valorem cigarette tax at the wholesale level. Data sources are: for road distances between largest metropolitan areas, Rand McNally *Road Atlas* (1972); for cigarette prices and taxes, Tobacco Tax Council (1972), for consumer price index, U.S. Bureau of Labor Statistics (various years).

[30] We would expect within states a tax-induced shift toward higher-priced cigarettes such as 100 mm and those from vending machines. As a result, the average price per package should have risen. Since weights are fixed in the data, the regression coefficient does not reflect this phenomenon.

[31] The cost of financing the tax in terms of extra interest can account for only a small fraction of the amount by which the coefficient differs from unity.

[32] The size of New Hampshire makes it unlikely that cigarettes would be tailor-made for it. The scope for removing characteristics is quite limited, then.

[33] In scatter diagrams for individual years the slope of the unit tax, fitted by free hand, is consistently higher than one, and New Hampshire has consistently the largest negative deviation.

VIII. Final Comments

If economists are set to offer policy prescriptions, the preceding analysis warns of a serious pitfall. Our model, and our results, cast doubt on the notion held by many economists with respect to second-best solutions—that taxes can be used to correct distortions—as we argue that a tax will have additional "side effects."

For instance, a gasoline tax is often regarded as a means to combat air pollution. Since the tax is actually constant per gallon, market substitution will occur toward high-octane gasoline which has become relatively cheaper But high octane ratings are produced by lead, now believed to be a major air pollutant, and the tax which results in a higher per-gallon use of lead may actually have increased the absolute level of pollution. The asserted purpose of the tax thus may have backfired [34]

While taxes do affect resource allocation, the common normative statement that they misallocate resources makes little sense If that is all they do, why would they ever be imposed? Until we can spell out the objectives of policymakers and the constraints they face we cannot attach meaning to terms such as misallocation or distortion. What we can do, however, is to determine the exact nature of any tax so that we can more accurately predict its effects.

Let us briefly summarize the thrust of the argument Commodities as transacted in the market are complex, and the margins with respect to which optimization takes place are numerous Because commodity tax statutes will not generally cover all these margins, any tax will induce multiple changes not only in resource allocation away from the taxed commodity and into others but also in the "quality" of the commodity and how it is transacted, a substitution away from the taxed attributes and into the others.

Adjustment in the face of incomplete specification of commodities implies upgrading of quality in the case of a unit tax and downgrading in that of an ad valorem tax. A test of the effect of taxes on cigarettes refutes the conventional hypothesis that quality will not change with a tax.[35] The results are rather consistent with our model, which claims that the effects of unit and of ad valorem taxes will differ both from each other and from those predicted by the conventional model

---

[34] Plott (1966, pp. 84–87) arrives at a similar result for a similar reason.

[35] The results for gasoline and liquor are also consistent with our model.

The two taxes produce different effects because of the different specification of what, exactly, is being taxed. Given the definition (and enforcement) the market will adjust with respect to each of the changeable characteristics. Although these changes are bound to be numerous and will differ depending on the type of tax, their extent is sharply constrained. They will be carried only to the point that minimizes the sum of the dollar value of the resultant inefficiencies and of the tax paid.

Appendix

Tests for Gasoline and Alcohol

Gasoline.—An attempt to test our hypothesis using data on gasoline taxes produced mixed results, vaguely encouraging but quite erratic. The main problems seem to be that (1) at any moment the wholesale price differs widely across markets, and (2) identification is hard to establish.

The data used are reported by the monthly magazine, *National Petroleum News* (1971), where for several years monthly observations are given on the wholesale price, tax, and retail price of regular gasoline in 55 cities.[36]

Regressions were run for 4 months in 1971 The dependent variable was retail price per gallon and the independent variables were wholesale price and tax. With unlagged independent variables the coefficient for tax is 0.78 (0.078 SE) and for wholesale price 1 18 (0.031 SE). The forces that make us expect a "large" coefficient for the tax variable apply to other cost components.[37] The coefficient for the wholesale price variable, at least, is comforting.

The variability in the wholesale price may well be due to differing demands among cities, while our hypothesis is in terms of shifts of cost schedules rather than shifts along them. Thus one cannot accept the above results at face value. When the two independent variables were lagged in an attempt to assure that the direction of causality in the data is from the independent to the dependent variable, the coefficient for the tax became 1.05 (0 23 SE) and that for the wholesale price 0 29 (0.09 SE).

It is clear that a more complete model and possibly better data are required to obtain satisfactory results for gasoline.

Alcohol.—The data used are from a secondary source.[38] A price breakdown is provided of 12 distilled spirits in New York State for 1971. These data enable us to run a regression in which the retail price per case is the dependent variable and tax per case and wholesale price (net of tax) are the independent variables Since measurement of one of the observations (imported brandy) presented problems, regressions were run with and without it. With the entire 12 observations the tax coefficient was 1.31 (0.40 SE) and that for the wholesale price 1.30 (0.15 SE). With 11 observations the respective coefficients were 1.51 (0.43 SE) and 1.26 (0.14 SE).

---

[36] "Regular" gasoline does not sufficiently define the commodity; octane ratings and other characteristics differ among regions and brands. I was unable to correct for these, however.

[37] Therefore, the two coefficients should be approximately equal, but here the difference between them is significant.

[38] Daniel Wisecarver (1974) kindly permitted me to use his data from an unpublished paper (tables 1 and 6).

The tax coefficients are larger than one, but not significantly so, and they are of the same order of magnitude as the coefficients for the wholesale price variable. The results, while not overwhelming, tend to confirm the hypothesis.[39]

References

Due, John F. *Government Finance An Economic Analysis*. Rev. ed. Homewood, Ill.: Irwin, 1959.

Gould, John P , and Segall, Joel. "The Substitution Effects of Transportation Costs " *J.P.E.* 77, no. 1 (January/February 1969) 130–37.

Griliches, Z. "Hedonic Price Indexes for Automobiles An Econometric Analysis of Quality Change " In *The Price Statistics of the Federal Government*, National Bureau of Economic Research. New York Nat Bur Econ. Res., 1961.

Harberger, A. C "Taxation, Resource Allocation, and Welfare." In *The Role of Direct and Indirect Taxes in the Federal Revenue System*, National Bureau of Economic Research and Brookings. Princeton, N.J Nat Bur Econ Res , 1964.

Lancaster, Kelvin J. "A New Approach to Consumer Theory " *J.P.E.* 74, no. 2 (April 1966): 132–57.

Lane, Frederic C. "Tonnages, Medieval and Modern " *Econ. Hist. Rev.* 17 (December 1964) · 213–33

Money Magazine (November 1972).

National Petroleum News (various issues, 1971).

Niskanen, William A., Jr "The Demand for Alcoholic Beverages." Ph.D. dissertation, Univ Chicago, 1963.

Plott, Charles R. "Externalities and Corrective Taxes." *Economica* 33 (February 1966) 84–87.

Rand McNally & Co. *Road Atlas*. 103d ed. San Francisco Rand McNally, 1972.

Tobacco Tax Council, Inc. *The Tax Burden on Tobacco Historical Compilation*. Vol 7 Richmond, Va. Tobacco Tax Council, 1972.

U.S. Bureau of Labor Statistics *Consumer Price Indexes* Washington Government Printing Office, relevant years.

Wisecarver, Daniel "The Comparative Resource Allocation Effects of Specific vs ad Valorem Taxation of Distilled Spirits." Unpublished paper, March 1974.

[39] In a study by Niskanen (1963) one of the findings is that in the taxation of alcoholic beverages consumers bear more than 100 percent of the tax This offers additional confirmation of the hypothesis

The Successes and Failures of Professor Smith

George J. Stigler

University of Chicago

The criterion employed to judge successes and failures has been the acceptance or nonacceptance of Smith's theory by his immediate successors. This judgment has been supplemented with a more personal one of scientific fruitfulness. He had one overwhelming and proper success—the theorem on resource allocation under competition—and several minor successes. He also had one modest success that was improper (i.e., unfortunate)—the distinction between productive and unproductive labor. Smith's most important analytical failure was the hierarchy of employments of capital. Three failures that should have been successes receive special attention: his wage theory, his rent theory, and the analysis of the division of labor. The reasons for the differing fates of these theories are examined.

The Wealth of Nations first appeared on March 9 of 1776, and perhaps sufficient time has now passed to permit a fair estimate of Professor Smith's triumphs and failures. It is a subject in which Smith himself would have displayed a vivid and natural interest. John Rae recounts his final days:

> When Smith felt his end to be approaching he evinced great anxiety to have all his papers destroyed except the few which he judged to be in a sufficiently finished state to deserve publication, and being apparently too feeble to undertake the task himself, he repeatedly begged his friends Black and Hutton to destroy them for him. . . . Black and Hutton always put off complying with Smith's entreaties in the hope of his recovering his health or perhaps changing his mind; but at length, a week before his death, he expressly sent for them, and asked them then and there to burn sixteen volumes of manuscript to which he directed them. This they did without knowing or asking what they contained. [1895, p. 434]

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved

Only a man acutely sensitive to the opinion which posterity would hold of him would insist upon such an act. Of course Smith was wrong: there is no amount of mischief and nonsense in 16 volumes which we would not have forgiven, especially since we know he was given to neither mischief nor nonsense

If the time is ripe, I am less certain of the qualifications of the writer There is a game I sometimes play with children; I call it "Three Questions " If all three questions are answered correctly I promise $1 million, no doubt the Securities and Exchange Commission will eventually prohibit the game, or the Federal Reserve System will make it viable The first two questions present no difficulty perhaps the number of brothers and sisters the child has, and the city in which it lives The third question is a different matter Once I asked, "Who was Adam Smith's best friend?" The reply from this child was, "You are, Uncle George " I had someone like David Hume or James Hutton or Joseph Black in mind Still, I have long been a good friend of Smith, though I have no right to claim priority in his circle I do not believe that my friendship will distort the judgments I shall propose.

The task I set, in any event, is not the uninteresting one of praising or blaming Smith The triumphs of any scholar are those of his doctrines which he persuades his contemporaries and successors to heed carefully When Ricardo or John Stuart Mill or Torrens adopted a theory of Smith's that does not necessarily mean that they accepted it without qualifications but that their work and thoughts were directed by the formulation of Smith Smith's failures were, correspondingly, those theories which his successors either ignored or rejected out of hand When Smith was wrong we would naturally expect able successors to ferret out the error, but we shall also discover that some of Smith's finest theories suffered the fate of neglect In any event, it is the judgment of the science that is decisive in judging a scholar's achievements

There is, I hope, an intrinsic interest in Smith's triumphs and failures simply because he was as great an economist as has ever lived There is also a broader significance to my query can we determine the characteristics of theories that help or hurt their reception?[1]

I. The Proper Successes

A success or triumph is a proposition in economics that becomes a part of the working system (the so-called paradigm) of contemporary and subsequent economists They accept and *use* the proposition, with heavy emphasis upon the word "use," or they reject and *dispute* the proposition, with heavy emphasis upon "dispute." In either event, their work is influenced by the successful proposition and, indeed, measures the success. So I repeat: a theoretical analysis is a success if it becomes a part of the living economics of successors, and the success is ascribable to Smith if his formulation governs the later use of the theory, whether he invented it or not. Hence I shall not attempt to determine Smith's debts to his predecessors; suffice to say they were large, but much smaller than our debts to him One can say of Smith what Newton said of himself "If I have seen farther, it is by standing on the shoulders of giants." That is appropriate to comparisons with predecessors but not to comparisons with contemporaries· they had the same shoulders to stand on

---

[1] I tried to answer this question in Stigler 1965 (pp. 66–155). The present approach is essentially independent although the answers overlap

Smith had one overwhelmingly important triumph. he put into the center of economics the systematic analysis of the behavior of individuals pursuing their self-interest under conditions of competition. This theory was the crown jewel of *The Wealth of Nations*, and it became, and remains to this day, the foundation of the theory of the allocation of resources. The proposition that resources seek their most profitable uses, so that in equilibrium the rates of return to a resource in various uses will be equal, is still the most important substantive proposition in all of economics.

I do not know whether to list as a second triumph one enormously successful application of this theory of competitive prices, namely, Smith's theory of the differentials in wage rates and profit rates among occupations The famous list of cost factors which would generate apparent but not real differences in rates of wages and profits—training, hardships, unemployment, and trust—were accepted, and in fact usually quoted verbatim, by Smith's successors for a century [2] This literature is the direct ancestor of Marshall's famous chapters on wages (1961, bk 6, chaps. 3–5), and of the modern theory of human capital So perhaps this special application of price theory deserves to be listed as his second success.

The third and final major success of Smith was his attack on mercantilism I measure a success by the impact of a scholar on other scholars, not his impact upon public thinking or public policy Smith's attack on protectionism in all its basic forms—tariffs, subsidies, compulsory use of domestic shippers, limitations on colonial enterprise, and the like—rested squarely on his theory of competitive prices The crucial argument for unfettered individual choice in public policy was the efficiency property of competition· the manufacturer or farmer or laborer or shipper who was seeking to maximize his own income would in the very process be putting resources where they were most productive to the nation· "Every individual is continually exerting himself to find out the most advantageous employment for whatever capital he can command. It is his own advantage, indeed, and not that of the society which he has in view. But the study of his own advantage naturally, or rather necessarily leads him to prefer that employment which is most advantageous to the society" (Smith 1976, 1:454). This application of price theory was again a corollary of the main proposition, but its development was so extensive and its success so great that it clearly deserves to be called Smith's third major triumph

---

[2] The role of trust was not analysed satisfactorily by Smith and its acceptance was much less complete (for a modern interpretation of it, see Becker and Stigler [1974]). Smith's fifth source of differences, the uncertainty of success, was not a cost-based differential, and it was much disputed rather than generally accepted.

I have been most parsimonious in quoting Smith, and it is well to present a sample of the power of his argument. Here is how he describes one section of the English policy of mercantilism

> To found a great empire for the sole purpose of raising up a people of customers, may at first sight appear a project fit only for a nation of shopkeepers It is, however, a project altogether unfit for a nation of shopkeepers, but extremely fit for a nation whose government is influenced by shopkeepers Such statesmen, and such statesmen only, are capable of fancying that they will find some advantage in employing the blood and treasure of their fellow-citizens, to found and maintain such an empire Say to a shopkeeper, Buy me a good estate, and I shall always buy my clothes at your shop, even though I should pay somewhat dearer than what I can have them for at other shops, and you will not find him very forward to embrace your proposal But should any other person buy you such an estate, the shopkeeper would be much obliged to your benefactor if he would enjoin you to buy all your clothes at his shop England purchased for some of her subjects, who found themselves uneasy at home, a great estate in a distant country The price, indeed, was very small, and instead of thirty years purchase, the ordinary price of land in the present times, it amounted to little more than the expence of the different equipments which made the first discovery, reconnoitred the coast, and took a fictitious possession of the country The land was good and of great extent, and the cultivators having plenty of good ground to work upon, and being for some time at liberty to sell their produce where they pleased, became in the course of little more than thirty or forty years (between 1620 and 1660) so numerous and thriving a people, that the shopkeepers and other traders of England wished to secure to themselves the monopoly of their custom. Without pretending, therefore, that they had paid any part either of the original purchase-money, or of the subsequen expence of improvement, they petitioned the parliament that th cultivators of America might for the future be confined to thei shop; first, for buying all the goods which they wanted fro

> Europe; and, secondly, for selling all such parts of their own produce as those traders might find it convenient to buy. [For they did not find it convenient to buy every part of it. Some parts of it imported into England might have interfered with some of the trades which they themselves carried on at home. Those particular parts of it, therefore, they were willing that the colonists should sell where they could, the farther off the better; and upon that account proposed that their market should be confined to the countries south of Cape Finisterre] A clause in the famous act of navigation established this truly shopkeeper proposal into a law [1976, 2 613–14]

From 1776 to today, the effect of this powerful attack, reinforced by the theoretical advances of Ricardo, Mill, and others, established a tradition of free international trade which even the most confirmed of economist-interventionists seldom feel equal to attacking frontally

There is a fourth considerable success to be credited to Smith. the formulation of the wages-fund theory This theory explained the short-run level of average wages by the ratio of funds for the payment of labor (the wages fund) to the number of laborers employed It was saved from being a tautology by the implicit condition that over moderate periods of time the wages fund was approximately constant in size Putting aside the question whether it was a useful theory (I have argued that it was [Stigler 1968]), there is no doubt that it dominated the next 100 years of English economics The uncertainty is how clearly Smith formulated the theory. He definitely asserted the essence of the theory, as when he says. "The demand of those who live on wages, it is evident, cannot increase but in proportion to the increase of the funds which are destined for the payment of wages These funds are of two kinds, first, the revenue which is over and above what is necessary for the maintenance, and, secondly, the stock which is over and above what is necessary for the employment of their masters" (1976, 1·86, also 1 110, 453). Smith's theory of wages will be shown below to rest on a wages-fund mechanism The only real misgiving is that Smith did not explicitly define the contents of the wages fund.[3]

I am painting with a wide brush insights and arguments of lesser scope, which would be a source of fierce pride to lesser economists, do not deserve inclusion here The famous paradox of value concerning diamonds and water, for example, which posed in inescapable form the central problem for the marginal utility theory, would deserve attention in any lesser-man's work. But the first three of these four successes I distinguish have become a permanent part of economics.

---

[3] This is Taussig's main reservation in the standard history of the wages fund (1899, pp. 150–51) But the fund is not so easily defined as Smith's successors and Taussig believed—it was not simply, or all of, "the consumable goods, in dealers' hands, ready for purchase by laborers" (ibid., p 148) The period of advance of the wages determines how "ready" the goods need be, as just one complication

II. The Improper Successes

An improper success is an error or an infertile and undevelopable subject or method of analysis—but one that is influential with contemporaries or successors Most demonstrable errors, one hopes and believes, are soon ferreted out, but the analysis that somehow fails to identify and organize and exploit a useful body of knowledge can only be discovered with time.

I would propose only one significant topic in Smith's work that meets this description· his theory of productive and unproductive labor.

> There is one sort of labour which adds to the value of the subject upon which it is bestowed· There is another which has no such effect. The former, as it produces a value, may be called productive, the latter, unproductive labour. [The] labour of the manufacturer fixes and realizes itself in some particular subject or vendible commodity, which lasts for some time at least after that labour is past It is, as it were, a certain quantity of labour stocked and stored up to be employed, if necessary, upon some other occasion The labour of the menial servant, on the contrary, does not fix or realize itself in any particular subject or vendible commodity His services generally perish in the very instant of their performance. [1976, 1 330]

The purpose of the distinction is clear: if we identify productive labor by the characteristic that its product can be accumulated, then capital formation can take place only out of the product of productive labor The difficulties with the distinction are two. Even if Smith is correct, the extensive employment of productive labor merely *permits* the accumulation of capital, and the actual formation of new capital requires a wholly independent act of saving Since most tangible products are not accumulated as capital but are currently consumed, there could be the loosest of connections between the share of labor that is productive and the rate of capital growth.

There is a second difficulty: there are investment acts which are not the result of productive labor. Investments in what we now call human capital do not become incorporated in a tangible, saleable commodity as commonly understood. Yet Smith agrees that one portion of the stock of a society consists of the acquired useful abilities of its inhabitants—to which he should have added the discovery of new knowledge: "The acquisition of [useful] talents, by the maintenance of the acquirer during his education, study, or apprenticeship, always costs a real expence, which is a capital fixed and realized, as it were, in his person. These talents, as they make a part of his fortune, so they do likewise of that of the society to which he belongs (1976, 1 282)." Unless we include instruction and training as productive labor—and Smith lists "men of letters of all kinds" as unproductive labor—the existence of productive labor is not even necessary to capital formation.[4]

The concept of productive labor never made a deep impression on Smith's successors. Senior and McCulloch denied the distinction and John Mill refined it almost out of existence [5] So it was a small improper success.[6]

III. The Proper Failures

Smith's failures to persuade economists were, like his successes, of two sorts failures that were proper, and failures that should have been successes We consider first the proper failures A proper failure contains an analytical error, or it presents an empirically trivial or mistaken view of the world.

The most conspicuous of Smith's proper failures was the hierarchy of employments of capital, presented in book 2, chapter 5, "Of the Different Employments of Capital" "A capital may be employed in four different ways· either, first, in procuring the rude produce annually required for the use and consumption of society; or secondly, in manufacturing and preparing that rude produce for immediate use and consumption, or, thirdly, in transporting either the rude or manufactured produce from the places where they abound to those where they are wanted; or lastly, in dividing particular portions of either into such small parcels as suit the occasional demands of those who want them" (1976, 1·361) Although all four activities are essential to one another or to "the general convenience of the society," capital is more productive—that is, sets more labor to work and augments more the annual produce of the society—if applied earlier in this sequence of operations than if applied later. The argument is simple: the capital of a retailer employs only himself and possibly a clerk—the remainder of the capital goes to purchase the goods he sells and therefore to replace the capitals of earlier stages At the other end, "no equal capital puts into motion a greater quantity of productive labour than that of the farmer" (ibid , p 363), for all of his capital goes to support labor, and in addition the fertility of nature is enlisted

---

[4] William Playfair pointed out, in his edition of the *Wealth of Nations*, that even the employment of menial servants might be a productive act " a cook, for (example) is a menial servant, but in a tavern he enriches his master just as much as any other journeyman, and on the other hand, a servant that spins or sews for the use of her master in a private family, is only acting as a menial servant, she is just supplying his wants, and contributing to his comforts in the same manner as when she lights the fire, or washes the apartments, yet she is a productive labourer by this definition" (Playfair 1805, 2·2, n).

[5] See Mill 1967, pp. 280 ff A less influential figure, David Buchanan, was a strong supporter of the distinction (Buchanan 1817, pp 131–36)

[6] Hla Myint would wish to make the concept of productive labor, and of economic growth more generally, central to Smith's work, and a wholly proper success (1948, pp. 20–24). I would assert the contrary that growth is not the only important path to economic welfare in Smith's system, and the concept of productive labor is not important to growth. But the modesty of the success of Smith's distinction is a matter of historical record

That Smith was in error is unequivocal He allowed a system of financing to conceal the facts of economic life. If the consumer, instead of paying the retailer for the corn, had paid the farmer for raising it, the millwright for grinding it, the ship's captain for transporting it, and the retailer for stocking it, then everyone's capital would have gone exclusively to the direct support of production, but nothing essential would have changed.

If Smith had really incorporated this error into his theoretical system, the effects would have been disastrous As one important example, the argument for private control over investment would have been damaged beyond repair But it remained a local blemish (repeated however once, in Smith 1976, 2·573), duly refuted by McCulloch and ignored by Senior and J S Mill (McCulloch 1825, pp 143 ff). Only Malthus gave it warm approval (1820, pp 30 ff).

This error is commonly, and no doubt correctly, attributed to the influence of the Physiocrats there is no such thing as a free trip to Paris But this is a history of the error, not an explanation for Smith's commission of it (Does the explanation lie in his antiluxury viewpoint?—reference will be made below to this attitude)

A related error, and one to which Smith attached greater importance if measured by the number of times it recurs in the *Wealth of Nations*, is the assignment of a hierarchy of social usefulness to domestic trade, foreign trade, and the carrying trade for foreign nations (1976, 1 368 ff., 454, 495–96, 2:600–604, 610–11, 628–30) The internal trade, he argues, by the act of buying Scottish manufactures, carrying them to London, selling them and buying English corn to return to Edinburgh, replaces two British capitals, whereas the foreign trade replaces only one British capital and the carrying trade none In addition, the returns of local trade are quicker than in distant trade At this level of discourse, Smith is surely mistaken If these various trades are yielding equal annual rates of return on capital, a shift from foreign to domestic trade would reduce aggregate national output (although the export of capital can of course affect wages) This error received no greater approval from Smith's successors (thus, Ricardo 1951, pp. 350–51).

A very different error, and possibly not an error at all, is Smith's measure of value—which came from the same source as that which may have led him to overvalue agriculture. Smith was acutely sensitive to the instability of monetary measures of value, and an appreciable fraction of the *Wealth of Nations* is devoted to the chronicle of currency debasement and inflation. He proposes as the ultimate measure of value the disutility of 1 hour of ordinary labor:

> Equal quantities of labour, at all times and places, may be said to be of equal value to the labourer. In his ordinary state of health, strength, and spirits, in the ordinary degree of his skill and dexterity, he must always lay down the same portion of his ease, his liberty, and his happiness The price which he pays must always be the same, whatever may be the quantity of goods which he receives in return for it . Labour alone, therefore, never varying in its own value, is alone the ultimate and real standard by which the value of all commodities can at all times and places be estimated and compared [1976, 1 50–51]

Smith's error, if indeed it is an error, is to assume that the psychological cost of performing 1 hour of labor is more stable, in its significance to a person, than the psychological pleasure from the consumption of some bundle of goods The instability of labor disutility arises from at least three circumstances (*a*) It varies with the conditions of technology—for example, the lifting of heavy weights has been almost eliminated in a modern society (*b*) It varies with the degree of training of the worker the disutility of acquiring labor skills must be added to that of performing the work, and this addition was already increasing secularly with the progressive division of labor (*c*) It varies with the hours of labor, and hence with income The corresponding view of a bundle of consumer goods yielding constant satisfaction as the unit of value is free of the second difficulty, possibly free of the first (depending how one views new commodities), but of course not free of the third

Smith's rejection of consumption in fixing on a measure of value is attributable to his belief that luxuries are frivolous and yield illusory pleasures that vanish in the act of realization This view is extensively argued in his *Theory of Moral Sentiments* (1797, pt 4, chap. 1) and receives adequate expression in the *Wealth of Nations* [7] That Smith should attribute to almost all economic actors an illusion that greater wealth yields greater satisfactions, an illusion that is perhaps never pierced, is one of his greatest idiosyncrasies

Smith's third error, and again perhaps we should label it a misdirection, is his monetary theory, as presented in volume 1, book 2, chapter 2, "On Money " Smith believes that there is a fixed demand for money in a society, in the special sense that only a certain quantity of money will circulate and excessive sums will be exported (if the money is gold or silver) or be presented for redemption in gold (if the money is bank notes) The theory is tenable as a first approximation if, as Smith assumes, the foreign exchanges are fixed and the paper currency is fully convertible: the theory then is implicitly a simple purchasing power parity theory.[8]

---

[7] "For a pair of diamond buckles perhaps, or for something as frivolous and useless, [the feudal lords] exchanged the maintenance, or what is the same thing, the price of the maintenance of a thousand men for a year, and with it the whole weight and authority which it could give them" (Smith 1976, 1·418–19)

The complaint at Smith's theory is not that it is formally erroneous but that it represents a retrogression from the generality and predictive power of the monetary theory in Hume's essays.

IV. The Improper Failures

There remain the successes that Smith should have achieved, but did not It will appear paradoxical that his immense prestige and vast powers of persuasion should have failed to obtain acceptance of ideas that were correct, profound, and fecund.

The first of these superior theories was a rejection of the subsistence theory of wages Smith, it will be recalled, gave four explicit reasons for believing wages were not generally at a subsistence level in Great Britain. (1) Summer wages exceed winter wages, but the cost of subsistence varies inversely (2) Subsistence varies substantially in cost from year to year, but some wages change very slowly (3) Subsistence varies substantially from place to place, but wages vary less by place (4) Variations over time and place in the cost of subsistence are often inverse to those of wages (1976, I.91–93). All of these proofs, particularly the first two, suffer from a concentration on short-run correspondences of wages and the cost of subsistence, but they carry considerable weight In addition Smith offers the powerful long-run example of the differences in real wages between England and the American colonies (ibid., pp. 87–88), an example whose persistence made it stronger with each passing year

Smith proposed an alternative theory, and one which was surely more valid than the subsistence theory as a predictor of wage rates. He proposed that "the" wage rate of (say) unskilled labor was given by wage rate $=$ subsistence level $+ \lambda(\Delta\ [\text{capital}]/\Delta\ [\text{time}])$, $(\lambda > 0)$, that is, that population lagged changes in capital, so ". . it is in the progressive state, while the society is advancing to the further acquisition, rather than when it has acquired its full complement of riches, that the condition of the labouring poor, of the great body of the people, seems to be the happiest and the most comfortable. It is hard in the stationary, and miserable in the declining state" (ibid., p 99). This wage theory, despite its great plausibility, was easily vanquished for a generation by Malthus's simple theory (which set $\lambda = 0$).

---

[8] Smith assumes the full convertibility of paper money (ibid, p. 329).

A second of Smith's theories took slightly more than a century to achieve currency—it was his theory of rent. He consistently treated the rent of land as it should be treated: any one use of land had to pay a rent, which was a cost of production, to draw the land away from other uses, whereas for all uses combined, rent was a residual. This theory is present in volume 1, book 1, chapter 11, "The Rent of Land," with hardly any ambiguity but with hardly any explicitness. Indeed, I used to suspect my own reading of it until I discovered an early and wholly concordant treatment by D. H. Buchanan (1929). It is difficult in retrospect to see how the many recognitions of the alternative cost theory received so little attention, as when Smith says, "As an acre of land, therefore, will produce a much smaller quantity of the one species of food [meat] than of the other [corn], the inferiority of the quantity must be compensated by the superiority of the price. If it is more than compensated, more corn land would be turned into pasture, and if it was not compensated, part of what was in pasture would be brought back into corn" (1976, 1: 165, 168, 175, etc.). John Stuart Mill gave this theory timid recognition, and Marshall refused to give it full credit. But unlike the other theories of Smith under discussion, the correct theory here is only partly explicit and was fragmented in presentation, so he rather than his successors deserve the larger blame for its neglect.

The last of Smith's regrettable failures is one for which he is overwhelmingly famous—the division of labor. How can it be that the famous opening chapters of his book, and the pin factory he gave immortality, can be considered a failure? Are they not cited as often as any passages in all economics? Indeed, over the generations they are.[9]

The failure is different: almost no one used or now uses the theory of division of labor, for the excellent reason that there is scarcely such a theory. The description of division of labor was much enlarged in Babbage's account of manufactures, and the phenomenon lies at the base of that part of Marshall's theory of external economies which attends to localization of industry. There are more praises and even mild use of Smith's theorem that the division of labor is limited by the extent of the market in essays by Allyn Young and myself, and Ronald Coase's work on the firm is clearly in the line of descent (Young 1928; Coase 1937; Stigler 1951). But there is no standard, operable theory to describe what Smith argued to be the mainspring of economic progress.

Smith gave the division of labor an immensely convincing presentation—it seems to me as persuasive a case for the power of specialization today as it appeared to Smith. Yet there is no evidence, so far as I know, of any serious advance in the theory of the subject since his time, and specialization is not an integral part of the modern theory of production, which may well be an explanation for the fact that the modern theory of economies of scale is little more than a set of alternative possibilities.

---

[9] Babbage reports the dramatic story of how M. Prony was enabled, by reading Smith on this subject, to plan the great mathematical tables of the French revolutionary government (1842, pp. 191–95).

V. The Recognition of Success and Failure

It is a general rule of scientific work that a scholar's successes and failures are judged by his contemporaries, and their judgment is accepted by later scholars. Allowing for the fact that Smith wrote when there were few even part-time economists, so we may perhaps treat the early nineteenth century as near contemporary, his experience confirms this rule. Certainly all of Smith's successes, proper and improper, were achieved within 50 years of their initial publication All of his failures, proper and improper, were similarly achieved in this period. It is almost (but not quite) tautological that proper successes and proper failures be promptly recognized at such, but it is not inevitable that an improper success (recall that Smith's on productive and unproductive labor was a modest success) will soon be dispatched

As for improper failures, the interesting point is that they do not influence the later adoption of the neglected contribution When the theory of rent was finally set straight—a development which took place primarily in the period from 1890 to 1910—the correct formulation owed nothing to Smith's profound but inarticulate insight When the Malthusian wage theory was abandoned, as it was on an ever-widening scale from 1825 on, it was not Smith's highly plausible theory of population lags that was adopted In fact, for a long time no alternative theory was adopted, and population receded from economic attention It is perhaps rash to complete the trilogy, for these are three scientific tragedies, by asserting that when a theory of specialization comes it will owe little to Smith, but on historical grounds it is even rasher to assert the converse

There are, of course, exceptions to the rule that much later scientific descendants accept the judgment of a man's contemporaries: perhaps Cantillon, Von Thünen, Cournot, and Gossen are the leading exceptions in the history of economics. These exceptions were for Schumpeter the source of considerable lament at the closed minds and narrow visions of contemporaries (Schumpeter 1954, pp 463 ff., and elsewhere), but it seems unfruitful to expect of a science that it immediately value all scientific work at its ultimate worth, never erring in deficit or in excess

The rule of the dominance of contemporary judgments, however, rests upon another basis: science is a *social* pursuit of knowledge,[4] not a census of independent individuals. A scholar who does not influence his contemporaries—who does not persuade them to work differently—is not an effective member of that science. Occasionally he may indeed have been too farsighted—inventing disk brakes before the internal combustion machine was known—but this is so rare a cause of failure that there cannot be many less efficient ways to discover good unexploited ideas than by reading earlier literature The overwhelming cause of failure of scholars is that their ideas were erroneous or infertile or their development too primitive to provide useful guidance to their contemporaries That is the proper reason for a judgment of failure in a social enterprise, and later scholars are quite sensible in accepting the verdict in perpetuity.

VI. The Sources of Success and Failure

In the long run nothing is more essential to a theory than that it be right, but we cannot even pause for a new sentence before remarking that rightness means limited wrongness The theory must help in explaining to the world that economics is attempting to understand, and a partial explanation is better than none

Logical error is sometimes enough to disqualify a theory—that is why Smith's four-layered hierarchy of employments of capital never had a prospect of scientific prosperity Usually, however, it is possible to retain the substance of a theory in a logical reformulation Perhaps that is the reason that logical criticisms of Smith's hierarchical theory were not common—economists like Ricardo simply ignored this theory, and it is difficult to doubt that their aversion was at least partly due to a belief that the whole approach was infertile and contrived

Blemishes, however, will exist in every theory the logic may be reasonably rigorous—although standards of rigor are not unchanging—but the very formulation of a problem will, in time, prove to be obtuse.

The acceptability of Smith's theories, logic aside, was very little influenced by the strength of the *specific* evidence he gave in their support. His strong empirical arguments against a subsistence theory of wages were ignored, whereas his support for the theorem on equalization of rates of return under competition was only casual and anecdotal. To say that a proof is nonspecific and nonquantitative is not to say that it is unweighty: if the theorem was congruent with widely observed phenomena—the growing number of members of prosperous trades, the fall of prices of new goods over time— then one was prepared to follow the theorem into unobserved places A century later, however, when Cliffe-Leslie denied the tendency of rates of return to approach equality, there was precious little documented evidence to refute him

But the proof of the ubiquity of the division of labor certainly also met this test of congruence with common observation: in fact, irresistibly so. Ask a modern economist to name an instance of *non*specialization of labor and he will be lucky to remember Robinson Crusoe. Yet, as we saw, there has been scarcely any systematic or regular use of this concept in economic analysis.

We have already hinted at the difference between the fates of the theorems on division of labor and on rates of return. The latter was a generalization of enormous power and could be used immediately on the most obtrusive and important questions: why some occupations earned more than others, why mercantilism and similar state interventions, as well as private monopoly, led to misallocations of resources, and who would bear various taxes. The theory of division of labor is not devoid of consequences—that it is limited by the extent of the market makes it relevant to protectionism, for example—but the uses were few and far between.

So Smith was successful where he deserved to be successful—above all in providing a theorem of almost unlimited power on the behavior of man. His construct of the self-interest-seeking individual in a competitive environment is Newtonian in its universality That we are today busily extending this construct into areas of economic and social behavior to which Smith himself gave only unsystematic study is tribute to both the grandeur and the durability of his achievement.

References

Babbage, C. *On the Economy of Machinery and Manufactures.* London: Knight, 1842.

Becker, G. S., and Stigler, G. J. "Law Enforcement, Malfeasance, and Compensation of Enforcers." *J Legal Studies* 3 (January 1974): 1–19

Buchanan, D *Observations on the Subjects Treated of in Dr. Smith's Inquiry into the Nature and the Cause of the Wealth of Nations.* London: Ogles, Duncan & Cochran, 1817

Buchanan, D H "The Historical Approach to Rent and Price Theory." *Economica* 9 (June 1929) 123–55. Reprinted in *Readings in the Theory of Income Distribution* Philadelphia. Blakiston, 1946.

Coase, R. "The Nature of the Firm." *Economica,* n.s 4 (November 1937): 386–405.

McCulloch, J. R. *The Principles of Political Economy.* Edinburgh: Tait, 1825.

Malthus, T. R. *Principles of Political Economy.* London: Murray, 1820.

Marshall, A. *Principles of Economics.* London: Macmillan, 1961.

Mill, J. S. *The Collected Works of John Stuart Mill.* Vol. 4, *Essays on Economics and Society.* Toronto: Univ. Toronto Press, 1967.

Myint, Hla. "The Welfare Significance of Productive Labour." *Rev. Econ. Studies* 9 (Winter 1943). 20–31.

Playfair, W., ed. *The Wealth of Nations by Adam Smith.* London: Cadell & Davies, 1805.

Rae, John. *Life of Adam Smith.* London: Macmillan, 1895.

Ricardo, D. *Works and Correspondence.* Vol. 1, *Principles of Political Economy.* Cambridge: Cambridge Univ. Press, 1951.

Schumpeter, J. A. *History of Economic Analysis.* New York: Oxford Univ. Press, 1954.

Smith, A. *The Theory of Moral Sentiments.* Dublin. Beatty & Jackson, 1797.

———. *The Wealth of Nations.* Oxford· Clarendon, 1976.

Stigler, G. J. "The Division of Labor Is Limited by the Extent of the Market." *J.P.E.* 59, no. 3 (June 1951): 185–94.

———. *Essays in the History of Economics* Chicago. Univ Chicago Press, 1965.

———. "Mill on Economics and Society." *Univ. Toronto Q.* 28 (October 1968): 96–102.

Taussig, F. W. *Wages and Capital.* New York. Appleton, 1899.

Young, A. A. "Increasing Returns and Economic Progress." *Econ. J.* 38 (December 1928) 527–42.

Futures Trading and Market Information

Charles C. Cox

Ohio State University

This paper investigates the effect of organized futures trading on information in spot markets. First, a model is developed that relates spot-price behavior and market information. The model can be viewed as a particular efficient markets model, this connection provides additional implications about price behavior and information. Next, price series for six different commodities are investigated for an information effect of futures trading For each commodity, the empirical evidence indicates that futures trading increases traders' information about forces affecting supply and demand.

The influence of futures trading on commodity prices has long been a controversial subject. For some 80 years, there have been farmers and other agricultural interests who have claimed that futures trading destabilizes spot prices and thereby imposes losses on producers and consumers. Congress has decided that futures trading frequently causes unreasonable price fluctuations and has enacted several laws regulating organized trading in commodity futures: trading in onion futures is prohibited, and futures trading in all other commodities is regulated by the Commodity Futures Trading Commission.[1] The congressional hearings on regulation of futures markets show that the regulators have neither a theory of destabilizing futures trading nor empirical evidence of destabilized prices. Several economists have studied futures trading and price variability by comparing price ranges or variances in a period of no

---

This paper is drawn from my doctoral dissertation ("The Regulation of Futures Trading," University of Chicago, 1975) I am indebted to George Stigler, Lester Telser, and the late Reuben Kessel for valuable advice and criticism.

[1] Trading in onion futures is prohibited by P.L. 85-839, August 28, 1958, 72 Stat. 1013. Trading in other commodities is regulated by the Commodity Futures Trading Commission Act of 1974, P.L. 93-463, 88 Stat. 1389 Prior to 1974, futures trading in most agricultural products was regulated by the Commodity Exchange Act.

[*Journal of Political Economy*, 1976, vol 84, no. 6]
© 1976 by The University of Chicago. All rights reserved.

futures trading and a period of active futures trading.[2] Some conclude that futures trading reduces price fluctuations; others find no difference. None of these studies offer much theoretical explanation of the empirical results.

This paper takes a different approach to the price effects of futures trading. I consider the relation between futures trading, market information, and spot prices, and I develop a model to analyze the information effect of futures trading. In the context of the model, empirical evidence on commodities' price behavior indicates that futures trading increases the information incorporated in a commodity's spot price. The results are inconsistent with the view that futures trading destabilizes spot prices in a way that is harmful to traders who handle the physical commodity.

I. Price Effects of Futures Trading

The relation between spot and futures prices together with certain characteristics of futures markets suggest that price effects of futures trading result from a change in the state of firms' price expectations. Telser (1958, 1967) shows that a futures price is an average of traders' expectations of the spot price that will prevail at the futures contract's maturity, and that for storable commodities, expected and current spot prices differ only by the net marginal cost of storage. Expected and current spot prices are determined simultaneously, and any change in expected prices induces a reallocation of the commodity between storage and current supply, thereby affecting the spot price. Futures trading can effect a state change in price expectations by altering the quantity of traders' information. Expected prices depend on current information about future supply and demand: more informed predictions are more accurate than less informed predictions. If information were free, all firms would have full knowledge of the evidence on future conditions and there would be no reason to believe that futures trading could influence price expectations. However, in real markets bits of information are dispersed among numerous individuals, the information changes frequently and is costly to acquire and communicate. Therefore, firms' price expectations reflect information that is neither complete nor perfectly accurate, and the introduction of futures trading can produce either more or less informed price expectations

---

[2] Hieronymus (1960), Working (1960), Gray (1963), and Johnson (1973) studied onions. Working, Hieronymus, and Gray concluded that futures trading stabilized onion prices, but Johnson concluded that there was no price effect of futures trading. Naik (1970) analyzed groundnut, linseed, and hessian. He concluded that futures trading reduced the variation in groundnut and linseed prices but did not affect hessian prices. Hooker (1901) and Tomek (1971) studied wheat, both concluded that the variation in wheat prices decreased in periods with futures trading. Emery (1896) studied cotton and concluded that futures trading reduced the yearly range of cotton prices. Taylor and Leuthold (1974) investigated cattle and Powers (1970) investigated pork bellies and cattle. Both studies concluded that short-term price variations are significantly reduced when there is futures trading.

There are at least two reasons for believing that futures trading can alter the amount of information reflected in expected prices. First, organized futures trading attracts an additional set of traders to a commodity's market: speculators who acquire and evaluate information in order to predict prices but who do not handle the physical commodity.[3] When these speculators have either a net long or short position in the futures market, hedgers (firms that deal in the physical commodity) have a corresponding net short or long position which causes the amount of stocks held for later consumption to be different than it would have been in the absence of futures trading. With different stocks, firms' price expectations are different too Because the speculators added by futures trading may be more or less informed about future conditions than traders who handle the physical commodity, the futures price can be consistently either a more or less accurate prediction of the spot price than the firms' expectations without futures trading. Critics of futures trading sometimes contend that speculators in futures markets are generally uninformed amateurs, while those more favorable to speculators emphasize that specialization should bring efficiency in price predictions. Relatively informed speculators would earn profits, while relatively uninformed speculators would find themselves suffering losses over time. Nevertheless, if uninformed speculators are willing to pay for the chance to earn profits from trading, and if new uninformed speculators enter the market and replace those who cannot sustain losses, a set of uninformed speculators would survive. A priori, it is impossible to determine whether speculators attracted by futures trading are more or less informed on average than other traders in the market Empirical evidence on this question is inconclusive: the results of several studies indicate that large-scale, professional speculators can profitably forecast commodity prices, but small traders cannot.[4]

---

[3] Speculators take a long or short position in the futures market when they expect the futures price to rise or fall Relatively low costs of transacting in futures markets make it worthwhile for these speculators to close out their positions with an offsetting sale or purchase of futures contracts rather than accepting delivery of and selling, or acquiring and delivering, the physical commodity. With futures trading, speculators can bear price risks whenever they expect profits without establishing trade connections for merchandising the commodity. This is not to say that there are no speculators trading a commodity in the absence of futures trading Whenever stocks of a commodity are held in an uncertain world, someone speculates by bearing the price risk of stockholding. Futures trading attracts additional speculators who would not trade the commodity without a futures market.

[4] Stewart (1949) analyzed futures-trading accounts that were mainly for small-scale speculators and found that losses greatly exceeded profits Houthakker (1957) and Rockwell (1967) used data on futures prices and traders' commitments and assumptions about trading to estimate gains and losses. Both found that large speculators earned profits and small speculators incurred losses. Using a similar method, Working (1931) estimated that speculators in wheat futures, as a group, incurred losses.

A second reason why futures trading can change the amount of traders' information concerns the cost of transacting. Without futures trading, individuals' expectations may differ widely, but often it is not worthwhile to communicate that information. A formal futures market reduces the cost of transacting because trading is completely centralized. Relative to dispersed trading and private negotiations, it is cheaper to identify potential traders, search for the best bid or offer, and negotiate a contract in a futures market. It becomes worthwhile for more individuals to trade and thereby communicate their information. The dispersed information on supply and demand is concentrated in one place and is all reflected in a single futures price.[5] All futures traders acquire this information, and because futures prices are widely publicized, the information incorporated in a futures price can be acquired cheaply by individuals who do not trade in futures markets. The magnitude of this effect depends upon the amount of the reduction in the cost of transacting. Organized futures trading should produce a larger effect the more decentralized a commodity's spot market is and the more numerous a commodity's traders are, since it is likely that the reduction in transacting cost will vary directly with these factors.

In order to investigate these effects of futures trading, it is necessary to specify the relations between information, price expectations, and the behavior of spot prices.

Market Information and Price Behavior

Consider a market with the following characteristics. Demand for the commodity fluctuates over time due to random shocks. Production is subject to a fixed lag, so the quantity produced depends on producers' price forecasts. Stocks of the commodity are held to smooth consumption over time. Information consists of a knowledge of the random shocks that affect demand, and, due to the cost of information, not all traders possess the most recent information. Expectations are formed as if the market calculates expected values conditional on the traders' information and the structure of the market. Although price expectations differ among firms—there is some distribution of price expectations—the model developed here does not take account of the dispersion of expectations. Firms' expectations are represented by a single expected price that is defined as an average of individual expected prices.[6] The following equations, based on a model by Muth (1961), describe the market and the effect on price behavior of a change in the quantity of traders' information.

---

[5] The role of a market price in summarizing and communicating information is discussed by Hayek (1945).

[6] Telser (1958) has used a market's expected price as an analytical device to develop an industry stockholding schedule. He suggests that the expected price can either be defined as an average or be derived from the relation between individual schedules and total market quantities. Houthakker (1968) has severely criticized the use of an expecte

$$C_t = \beta_0 - \beta P_t - \varepsilon_t \qquad \text{(Consumption demand)}, \tag{1a}$$

$$S_t = \gamma_0 + \gamma P_t^e \qquad \text{(Production)}, \tag{1b}$$

$$I_t = \alpha_0 + \alpha(P_{t+1}^e - P_t) \qquad \text{(Supply of storage)}, \tag{1c}$$

$$C_t + I_t = S_t + I_{t-1} \qquad \text{(Market equilibrium)}, \tag{1d}$$

where P_t is the price in period t, P_t^e is the market's expectation of the price that will prevail in period t, given the traders' information through period $t - 1$, and ε_t is a random disturbance affecting demand, C_t.[7] Assume that the ε's are independent and identically distributed, and that expected prices can be expressed as linear combinations of these disturbances The current price, then, is a linear function of the ε's,

$$P_t = \sum_{i=0}^{\infty} V_i \varepsilon_{t-i} + K. \tag{2}$$

From the expectations assumption, the expected price is the expected value of P_t conditional on the traders' information through period $t - 1$.

$$\begin{aligned} P_t^e &= V_0 E(\varepsilon_t) + (1 - f)V_1 E(\varepsilon_{t-1}) + fV_1\varepsilon_{t-1} + \sum_{i=2}^{\infty} V_i\varepsilon_{t-i} + K \\ &= fV_1\varepsilon_{t-1} + \sum_{i=2}^{\infty} V_i\varepsilon_{t-i} + K, \end{aligned} \tag{3}$$

where f is the fraction of the traders that possesses the information ε_{t-1} in period $t - 1$

In investigating a market for the effects of a change in the quantity of market information, that is, a change in f, neither the price expectations nor the random shocks can be observed directly The actual prices,

---

price for analysis of futures trading One of his major objections is that the procedure by which individual expected prices should be aggregated has not been specified. Furthermore, he objects to focusing exclusively on price expectations, as I do in this essay, because he argues that individual expected prices have ambiguous effects on traders' behavior unless a number of other variables, e g , the wealth effect of price changes, are taken into account. Clearly, there are variables other than price expectations that influence price movements, and an average of expected prices is only an approximation, but my goal is to concentrate on information and the behavior of spot prices. Whether my model is useful for analyzing this problem depends on its ability to predict actual price behavior

[7] In this model, random disturbances affect only demand because that is sufficient to make P_t a stochastic variable. Although it would seem more realistic to include a random term in production too, adding a disturbance like ε_t to (1b) neither increases nor changes the model's implications about information and price behavior. Under different assumptions about the disturbances, however, this conclusion does not hold. See Nelson (1975).

however, are observable. Therefore, to make this model operational it is necessary to solve for the expected price in terms of past prices. That is,

$$P_t^e = \sum_{j=1}^{\infty} Z_j P_{t-j} + H. \tag{4}$$

With equations (1) and (4), it is possible to derive a testable hypothesis about the effect of futures trading on the behavior of market prices. Hence, the task is to solve for the Z's of (4) in terms of the parameters of equations (1), then to find the effect of a change in f on the relation between prices. To do so, it is convenient to start by specifying the relationship between the V's implied by equations (1)–(3).

Substituting (2) and (3) into (1) yields the transformed market equilibrium equation:

$$\begin{aligned}[(\beta_0 + \alpha_0) - \beta K] - (\beta + \alpha) \sum_{i=0}^{\infty} V_i \varepsilon_{t-i} + \alpha \left(f V_1 \varepsilon_t + \sum_{i=2}^{\infty} V_i \varepsilon_{t+1-i} \right) \\ = [(\gamma_0 + \alpha_0) + \gamma K] + (\gamma + \alpha) \left(f V_1 \varepsilon_{t-1} + \sum_{i=2}^{\infty} V_i \varepsilon_{t-i} \right) \\ - \alpha \sum_{i=0}^{\infty} V_i \varepsilon_{t-1-i} + \varepsilon_t \end{aligned} \tag{5}$$

For (5) to hold for all possible shocks, the coefficients of the corresponding ε's must be equal, and the constant terms in brackets must be the same. Hence,

$$-(\alpha + \beta) V_0 + \alpha f V_1 = 1 \tag{6a}$$

$$\alpha V_0 - [\alpha(1 + f) + \beta + \gamma f] V_1 + \alpha V_2 = 0 \tag{6b}$$

$$\alpha V_{i-1} - (2\alpha + \beta + \gamma) V_i + \alpha V_{i+1} = 0 \quad (i = 2, 3, 4, \ldots). \tag{6c}$$

Equation (6c) is a homogeneous, second-order difference equation for which the solution is.

$$V_i = r^{i-1} V_1 \qquad (i = 2, 3, 4, \ldots), \tag{7}$$

where r is the smaller root $(0 < r < 1)$ of the characteristic equation for (6c) and is a function of α, β, and γ.[8]

Substituting (7) into (6b) and transforming yields

$$\phi \equiv \frac{V_1}{V_0} = \frac{\alpha}{[\alpha(1 + f - r) + \beta + \gamma f]}. \tag{8}$$

[8] Sufficient conditions for real and distinct roots in the characteristic equation for (6c are $\alpha > 0$ and $(\beta + \gamma) > 0$. The roots occur in reciprocal pairs, so only one root is les than one in absolute value. From (2), it follows that (6c) is stable or P_t would be infinite That is, the requirement that all bounded sequences of disturbances produce onl bounded sequences of prices implies that (6c) is stable. Therefore, the coefficient of th larger root in the general solution must be zero.

Now the solution of (4) follows easily; substituting (2) and (3) into (4) yields

$$fV_1\varepsilon_{t-1} + \sum_{i=2}^{\infty} V_i\varepsilon_{t-i} = \sum_{i=1}^{\infty}\left(\sum_{j=1}^{i} Z_j V_{i-j}\right)\varepsilon_{t-i} + K\left(\sum_{j=1}^{\infty} Z_j - 1\right) + H. \tag{9}$$

Like (5), (9) must hold for all possible shocks, so the corresponding coefficients of the ε's are equal.

$$fV_1 = Z_1V_0 \tag{10a}$$

$$V_i = \sum_{j=1}^{i} Z_jV_{i-j} \qquad (i = 2, 3, 4, \ldots). \tag{10b}$$

Substituting (7), (8), into (10) and solving for Z_i gives

$$Z_1 = f\phi \tag{11a}$$

$$Z_i = (r - f\phi)\phi(r - \phi)^{i-2} \qquad (i = 2, 3, 4, \ldots). \tag{11b}$$

Equations (11) produce the desired form for (4).

$$P_t^e = f\phi P_{t-1} + \sum_{j=2}^{\infty} [(r - f\phi)\phi(r - \phi)^{j-2}]P_{t-j} + H \tag{12}$$

Combining (12) and (1), the market equilibrium equation yields

$$\begin{aligned} P_t = {} & \frac{(\gamma_0 - \beta_0 + \gamma H)}{(\alpha f\phi - \alpha - \beta)} + \phi P_{t-1} \\ & + \frac{(r - f\phi)\phi[(\alpha + \gamma) - \alpha(r - \phi)]}{(\alpha f\phi - \alpha - \beta)}\sum_{j=2}^{\infty}(r - \phi)^{j-2}P_{t-j} \\ & + \frac{\varepsilon_t}{(\alpha f\phi - \alpha - \beta)} \end{aligned} \tag{13}$$

Let

$$b_0 = \frac{(\gamma_0 - \beta_0 + \gamma H)}{(\alpha f\phi - \alpha - \beta)} \tag{14a}$$

$$b_1 = \phi \tag{14b}$$

$$b_j = \frac{(r - f\phi)\phi[(\alpha + \gamma) - \alpha(r - \phi)](r - \phi)^{j-2}}{(\alpha f\phi - \alpha - \beta)} \qquad (j = 2, 3, 4, \ldots) \tag{14c}$$

$$u_t = \frac{\varepsilon_t}{(\alpha f\phi - \alpha - \beta)} \tag{14d}$$

then

$$P_t = b_0 + \sum_{j=1}^{\infty} b_jP_{t-j} + u_t. \tag{15}$$

Analysis of equations (14) shows that the b_j's have the following properties:

$$b_1 > 0$$

$$b_j \begin{cases} >0 \text{ if } j \text{ is odd} \\ <0 \text{ if } j \text{ is even} \end{cases} \quad \text{for} \quad 0 < f < 1 \quad \text{and} \quad (j = 2, 3, 4, \ldots)$$

$$|b_j| > |b_{j+1}| \qquad (j = 1, 2, 3, \ldots)$$

$$|b_j| < 1 \qquad (j = 2, 3, 4, \ldots).$$

Thus the model implies that when some traders have not acquired the most recent market information, the current market price will equal a linear combination of past prices plus a random-error term, that is, equation (15) The coefficients of past prices alternate in sign starting with a positive coefficient for the price immediately preceding the current price and the earlier the price, the smaller in absolute value is its coefficient.

An increase in market information acquired by the traders is equivalent to an increase in f The effect of increased information on the relationship between the prices is found by differentiating b_j with respect to f By equations (14):

$$\frac{\partial |b_j|}{\partial f} < 0 \qquad (j = 1, 2, 3, \ldots) \qquad (16a)$$

With an increase in market information, the coefficients of past prices in equation (15) all decrease in absolute value Also, it follows from equations (14d) and (15) that an increase in information decreases the variance of the price-forecast error.

$$\frac{\partial \sigma_u^2}{\partial f} < 0 \qquad (16b)$$

In the extreme case where all traders know the latest market information, that is, in the case where $f = 1$, equation (8) shows that $\phi = r$. Therefore, by (14)

$$b_1 = r, \qquad (17a)$$

$$b_j = 0 \qquad (j = 2, 3, 4, \ldots), \qquad f = 1. \qquad (17b)$$

Equation (15) then becomes

$$P_t = b_0 + rP_{t-1} + u_t. \qquad (18)$$

Why are price effects of uninformed traders not eliminated by traders with complete information? Put another way, why would fully informed traders limit their market positions instead of trading so that price is pegged at the level consistent with the unbiased estimate of the futur price? One reason is related to the idea that individuals' expectations can differ for the same information set, the theory requires only that a weighted average of individual expectations based on complete information equal the expected value of the future price Also, the fully informed traders need not be the same individuals in every period Consequently, lack of confidence in his price forecast would limit the commitment made by each informed trader even though the completely informed traders will, as a group, earn profits at the expense of the group of traders with incomplete information. If there is no risk aversion to limit the commitments of fully informed traders, then price effects of uninformed traders would persist only to the extent that costs of transacting make it unprofitable to completely eliminate them

Market Efficiency

The preceding implications about information and price behavior are compatible with theory developed in the efficient-markets literature—the model itself can be viewed as a particular efficient-markets model [9] The work on efficient markets assumes that market equilibrium can be stated in terms of expected values of price changes, and that prices in an efficient market fully reflect available information From these assumptions, it follows that the expected value of P_t conditional on all information at $t - 1$ is an unbiased estimate of P_t That is,

$$E[P_t - E(P_t \mid \text{all information at } t - 1)] = 0 \qquad (19)$$

For the model developed above, the assumption that market expectations equal conditional expected values allows equilibrium to be expressed in terms of expected values, available information is fully reflected in the price when all traders are informed. So price in an efficient market is represented by (18), from which the "fair game" property (19) follows directly.

$$E\{P_t - E[P_t \mid (\varepsilon_{t-1})]\} = E(u_t) = 0. \qquad (20)$$

Implications about the sequence of observed prices in an efficient market depend upon the stochastic processes generating price changes. Several efficient-markets studies assume that successive price changes are independent, identically distributed random variables, which implies that the sequence of price changes is a random walk. In my model, however, both systematic and random factors generate price changes Systematic changes are due to storage costs, and random changes are due to the disturbances that affect demand Hence, successive price changes are not independent, and the sequence of price changes is not a random walk even when current information is fully reflected in the market price. To see this consider first differences of (18):

---

[9] The efficient-markets literature is summarized in Fama (1970).

$$P_t - P_{t-1} = r(P_{t-1} - P_{t-2}) - u_{t-1} + u_t. \qquad (21)$$

Equation (21) shows that successive price changes are correlated, although the history of earlier price changes $(P_{t-2} - P_{t-3}, P_{t-3} - P_{t-4}, \ldots)$ adds no additional information about $P_t - P_{t-1}$. If there were no storage in the model, only the independent, identically distributed disturbances in demand (ε_t) would cause price changes and the sequence of price changes would be a random walk.

If information is not fully reflected in the price, efficient-markets theory implies only that the fair-game property will not hold Other than this, price behavior depends on the particular price-formation process and the information that does influence price The model in the present work focuses on differences in price behavior when more or less information is reflected in price, so it specifies both price formation and information in order to show how the relation between prices depends on the quantity of information (15). Since (15) is just one of many possible ways that a price series can behave, empirical analysis is necessary to determine if it corresponds to actual price behavior. However, (15) is consistent with efficient-markets theory in that the sequence of differences between observed and expected price is not a fair game When P_t is given by (15), expected value conditional on *all* information at $t - 1$ is the expected value of (18); hence

$$E\{P_t - E[P_t \mid (\varepsilon_{t-1})]\} = (\phi - r)P_{t-1} + \sum_{j=2}^{\infty} b_j P_{t-j} + \text{constant}. \qquad (22)$$

The history of past prices can be used to make a price forecast that is on average more accurate than the conditional expected value. As a result, when there are uninformed traders in the market, it may be possible for a chart reader to devise a profitable trading rule based only on past spot prices

The preceding model provides a framework for testing the hypothesis that commodity futures trading affects spot prices by increasing the market's information about forces influencing supply and demand. The coefficients of past prices in (15) reflect the quantity of the market's information, and since (15) is an autoregressive process, the coefficients can be estimated empirically from a time series of a commodity's prices. If the estimated coefficients are consistent with the implications of the model, and if there is a difference between the coefficients for periods with and without futures trading, the change can be interpreted as the contribution of futures trading to market information. Furthermore, an increase in information should decrease the returns from a trading rule. This too can be tested with a series of spot prices. The next section investigates the price behavior of six different commodities.

II. Empirical Analysis

Evidence on price effects of futures trading is presented here for onions, potatoes, pork bellies, hogs, cattle, and frozen concentrated orange juice (FCOJ). For each of these commodities, price effects of futures trading are estimated by comparing price behavior in a period with no futures trading to price behavior in a period when the commodity is traded in a formal futures market.

The most important conclusions of this empirical analysis are: (1) futures trading in a commodity increases the quantity of traders' information, (2) a spot market is more efficient in the sense that price more fully reflects available market information when there is futures trading, (3) the behavior of prices does not support the claim that producers and consumers are harmed by price effects of futures trading.

Data and Periods Examined

Price series for each commodity are compiled from trade journals that report wholesale spot prices[10] The prices are observed on 1 day at weekly intervals in the market where futures trading occurs and are for commodities with the same specifications as the basis grade for futures trading.

Sample periods depend on the length of time that a continual price series has been reported and on the date when futures trading in a commodity began. Periods when a commodity's prices were publicly controlled are also recognized in choosing the sample periods, since reported and transaction prices often differ during times of price controls.[11] Consequently, the lengths of the sample periods vary considerably; the longest sample contains 856 observations and the shortest contains 220 observations. The commodities, markets, and periods examined are listed in table 1

[10] Onion prices are taken from two basic sources. for the period 1928–59, prices are from the *Chicago Packer*, a trade journal for the Chicago wholesale vegetable market; for the period 1960–71, prices are from *Chicago Fresh Fruit and Vegetable Wholesale Market Prices*, a publication of the U.S. Department of Agriculture Two data sources are used because the *Packer* frequently failed to report onion prices after 1959, and the weekly price reports made available by the USDA were only for the years 1960 and later. Potato prices are all from the *Journal of Commerce* All prices for pork bellies, hogs, and cattle are from the *National Provisioner*. Frozen concentrated orange juice prices are taken from the *Journal of Commerce*.

[11] The series for onions and potatoes exclude the period of World War II price controls because reported prices were constant at ceiling levels and were not transaction prices. The other commodities' series start after World War II controls ended. Potatoes, pork, and cattle were subject to price controls for different periods during the Korean war, but prices for these periods are included in the samples for potatoes, pork bellies, hogs, and cattle because the periods of price control were relatively short and the data indicate that ceilings were not effective for most of those periods. After August 1971, various price-control programs at different times set ceilings on prices of all the commodities or their processed forms. Therefore, all of the price series end prior to, or at the start of, controls in 1971.

TABLE 1

COMMODITIES, MARKETS, AND SAMPLE PERIODS

Commodity	Market	Futures Trading Started	Sample Period with No Futures Trading	Sample Period with Futures Trading
Yellow onions	Chicago	Sept. 8, 1942,* terminated Nov 6, 1959†	Sept. 1928–Apr. 1942 Nov. 1959–Apr 1971	Sept. 1948–Oct. 1959
Maine potatoes	New York	Dec 2, 1941‡	Oct 1925–July 1941	Oct 1947–July 1971
Pork bellies	Chicago	Sept 19, 1961	Oct 1949–Sept 1961	Oct 1961–Sept. 1971
Hogs	Chicago§	Feb 28, 1966	Oct. 1949–Feb 1966	Mar. 1966–May 1970
Cattle	Chicago‖	Nov 30, 1964	May 1949–Nov 1964	Dec. 1964–July 1971
Frozen concentrated orange juice (FCOJ)	New York	Oct 26, 1966	Jan 1957–Oct. 1966	Nov 1966–Aug. 1971

* Onion and potato prices were subject to U.S government price controls from November 7, 1942 to September 6, 1946

† Public Law 85-839 prohibiting trading in onion futures became effective November 6, 1959

‡ Futures trading in potatoes was suspended during World War II and reopened January 17, 1946

§ Trading in live hogs at the Union Stockyards in Chicago stopped in May 1970

‖ Trading in live cattle at the Union Stockyards in Chicago stopped in July 1971

The Econometric Model

To investigate the impact of organized futures trading on traders' information an autoregression of the following form is estimated for periods with and without futures trading:

$$P_t = b_0 + \sum_{j=1}^{n} b_j P_{t-j} + u_t, \tag{23}$$

where P_t is the spot price at time t, u_t is a random term, j is an interval of 1 week. The expected values of the coefficients in (23) are

$$|b_j| > |b_{j+1}|,$$

$$|b_j| < 1 \qquad (j = 2, \quad , n)$$

$$b_j \begin{cases} >0 \text{ if } j \text{ is odd} \\ <0 \text{ if } j \text{ is even,} \end{cases}$$

for the fraction of fully informed traders $f < 1$ For $f = 1$, $0 < b_1 < 1$ and $b_j = 0$ $(j = 2, \quad , n)$. A priori, the value of b_0 is not determined. Equations (16) show that for a change in the state of traders' information

$$\frac{\partial |b_j|}{\partial f} < 0 \qquad \text{and} \qquad \frac{\partial \sigma_u^2}{\partial f} < 0.$$

An increase in the fraction of traders with a knowledge of the current market information decreases both the coefficients of past prices and the mean-square error of estimate Therefore, effects of futures trading on market information can be analyzed by comparing estimates of (23) for periods with and without futures trading

The theory in Section I shows price as an infinite-order autoregression (15), so the number of lagged prices in the econometric model (23) is determined on the empirical basis of minimum mean-square error of estimate. The best-fitting regressions for the sample periods with no futures trading are all obtained with five to 10 lagged prices. Autoregressions of the same order are then estimated for the periods with futures trading in order to compare parameters with and without futures trading.

The procedure used to estimate (23) is ordinary least squares, which yields consistent estimates of autoregressive parameters if the time series is stationary with independent and identically distributed disturbance terms. The price series investigated here are weakly stationary—they all exhibit seasonal patterns, but for periods spanning several years the prices vary about a fixed mean like a stationary series. Also, the sample correlograms damp out without peaks at the seasonal lags. Hence, the series are treated as stationary and no adjustments are made for seasonality. Analysis of residuals suggests that the disturbances are well behaved.

Only one of the regressions (FCOJ, 1957–66) has significant autocorrelation in the residuals.[12]

The price series used here are all nominal prices. This is dictated by the fact that there is no weekly price index for the entire sample period that can be used to transform the series into real prices. Deflating the prices with the monthly BLS Wholesale Price Index produces only minor changes in the regression results that are reported below But this may result because the procedure does not change the relation between prices that are observed in the same month.

The Effect of Futures Trading on Market Information

The regressions in table 2 estimate (23) and compare spot-price behavior in the absence of futures trading to spot-price behavior during periods of futures trading. The evidence strongly supports the hypothesis that futures trading increases the quantity of traders' market information.

First, it is reasonable to interpret differences in the regression estimates as changes in information because the model is able to predict relations between prices Estimated coefficients for each of the regressions conform to the expected values. For each of the 13 regressions, the absolute values of the coefficients b_2 through b_n are less than one and tend to conform to the prediction that $|b_j| > |b_{j+1}|$ Only one equation fits this pattern perfectly, but there are only three regressions for which half or more of the coefficients do not fit this pattern As expected, the coefficients tend to alternate in sign, and there is a tolerable correspondence between predicted and actual signs The regressions have, on average, 71 percent of the expected runs of signs For one regression all of the signs are the same as predicted, and only three regressions have as many as four coefficients that differ from the predicted signs The interval of 1 week between price observations is arbitrary, and this interval influences the actual patterns of coefficients The agreement between the estimated and expected patterns of coefficients, therefore, suggests that the model is useful for analyzing week-to-week price behavior.

Second, for every commodity comparisons of the regression estimates with and without futures trading indicate more informed traders during periods of futures trading As predicted for an increase in market information, the estimated coefficients during periods of futures trading are generally less in absolute value than the same coefficients in periods without futures trading. Thirty-five of the 48 coefficients decrease in absolute value, and only one of seven comparisons has an increase in as many as half of the coefficients.[13] However, the results are striking when significance of the coefficients is considered In the absence of futures trading, all of the commodities have one or more of the coefficients b_2 through b_n that are significantly different from zero at the 5 percent level For the regressions during periods of futures trading, only two of the 37 coefficients b_2 through b_n are different from zero at the 10 percent level! That is, the regression estimates not only show increased information during periods of futures trading for each commodity, for four of the commodities the estimates are consistent with all traders knowing the latest market information

---

[12] Residuals for each regression are tested for autocorrelation according to the procedure developed for autoregressions by Box and Pierce (1970). The Durbin-Watson statistic which is usually used to test for autocorrelated disturbances is biased toward accepting the null hypothesis of independent disturbances for any autoregression.

Put another way, increased market information reduces the relation between current and past prices When all traders know the latest information, the current price "depends" only on the immediately preceding price. Therefore, the evidence reflected in comparisons of regression coefficients with and without futures trading can be conveniently summarized by testing the joint influence of P_{t-2} through P_{t-n} on P_t The appropriate analysis-of-variance test is equivalent to testing the null hypothesis that $b_2 = b_3 = b_4 = \quad = b_n = 0$ Table 3 reports this test for each regression. The results add evidence that futures trading increases market information In the absence of futures trading, the set of lagged prices P_{t-2} through P_{t-n} is significant at high levels for every commodity For periods of futures trading, all the sets of lagged prices are insignificant even at low critical levels when the regression coefficients have low t-ratios So there is a consistent and significant decrease in the influence of past prices on the current price

The results for onions are important because trading in onion futures is prohibited, and it is possible to analyze onion prices both prior to and after the time of futures trading In both of the periods with no futures trading, 1928–42 and 1959–71, the current price depends significantly on lagged prices other than P_{t-1} as expected for a market with uninformed traders But during the period of futures trading, 1948–59, lagged prices other than P_{t-1} are insignificant as predicted for a market with fully informed traders Besides indicating increased information during the time of futures trading, this finding provides evidence that the information is due to futures trading rather than some other force that increases information over time. The same difference in significance of lagged prices is exhibited by pork bellies, hogs, and FCOJ when the periods prior to futures trading are compared to periods with futures trading. On the other hand, potatoes and cattle are noteworthy because the regression results indicate increased information with futures trading, but even then

---

[13] Although two-thirds of the coefficients decrease when there is futures trading, the set of coefficients $b_0, b_1, \quad , b_n$ differs significantly (at the 5 percent level) from the set when there is no futures trading for only two of the seven tests The significantly different regressions are for onions prior to futures trading and potatoes

TABLE 2

SPOT-PRICE REGRESSIONS

$$\left(P_t = b_0 + \sum_{j=1}^{n} b_j P_{t-j} + u_t\right)$$

Commodity and Period	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	R^2	SE of Estimate	Sample Size
Onions:														
Sept. 1928–Apr. 1942	0.051	0.657	0.257	−0.139	0.093	0.075						.794	0.249	473
	(2.281)	(13.746)	(4.355)	(−2.335)	(1.568)	(1.524)								
Sept. 1948–Oct. 1959	0.131	0.963	−0.065	0.021	−0.062	0.063						.844	0.358	369
	(3.147)	(18.377)	(−0.898)	(0.295)	(−0.850)	(1.205)								
Nov. 1959–Apr. 1971	0.131	1.018	0.018	−0.194	0.144	−0.052						.889	0.259	382
	(3.366)	(19.706)	(0.250)	(−2.728)	(2.018)	(−1.019)								
Potatoes:														
Oct. 1925–July 1941	0.066	0.893	0.148	−0.077	−0.155	0.102	0.049					.918	0.256	618
	(2.850)	(22.101)	(2.744)	(−1.441)	(−2.894)	(1.906)	(1.235)							
Oct. 1947–July 1971	0.267	0.872	0.103	0.0004	−0.073	0.023	−0.005					.849	0.414	809
	(5.093)	(24.685)	(2.196)	(0.009)	(−1.560)	(0.491)	(−0.131)							

TABLE 2 (*Continued*)

Pork bellies														
Oct. 1949–Sept. 1961	0.804	1.013	0.003	−0.148	0.190	−0.085						951	1.728	621
	(2.844)	(25.213)	(0.052)	(−2.524)	(3.244)	(−2.074)								
Oct. 1961–Sept. 1971	0.678	0.968	−0.038	0.037	0.050	−0.038						956	1.708	517
	(2.179)	(21.865)	(−0.622)	(0.596)	(0.810)	(−0.852)								
Hogs:														
Oct 1949–Feb. 1966	0.413	1.005	0.013	−0.001	−0.046	0.076	−0.057	0.090	−0.102			965	0.667	848
	(3.171)	(29.267)	(0.264)	(−0.021)	(−0.944)	(1.574)	(−1.178)	(1.842)	(−2.936)					
Mar. 1966–May 1970	0.672	0.954	0.096	−0.020	−0.084	0.060	−0.081	0.065	−0.020			945	0.753	212
	(1.705)	(13.617)	(0.996)	(−0.204)	(−0.877)	(0.624)	(−0.841)	(0.674)	(−0.282)					
Cattle														
May 1949–Nov 1964	0.358	0.791	0.116	0.176	−0.140	0.132	−0.031	0.034	−0.040	−0.052		973	0.704	804
	(2.315)	(22.344)	(2.568)	(3.880)	(−3.064)	(2.895)	(−0.675)	(0.751)	(−0.893)	(−1.472)				
Dec 1964–July 1971	0.623	0.787	0.235	0.057	−0.057	0.014	0.004	−0.098	−0.039	0.074		954	0.552	339
	(1.828)	(14.295)	(3.351)	(0.807)	(−0.797)	(0.201)	(0.056)	(−1.372)	(−0.547)	(1.329)				
FCOJ.														
Jan. 1957–Oct. 1966	0.754	0.930	0.070	0.027	0.100	−0.204	0.103	0.007	−0.049	0.128	−0.131	976	1.772	502
	(2.563)	(20.793)	(1.143)	(0.443)	(1.639)	(−3.333)	(1.691)	(0.113)	(−0.803)	(2.101)	(−2.957)			
Nov. 1966–Aug. 1971	0.706	1.013	−0.030	0.076	−0.072	0.001	−0.002	0.006	0.044	−0.054	−0.003	965	1.224	242
	(1.618)	(15.396)	(−0.321)	(0.811)	(−0.769)	(0.012)	(−0.026)	(0.061)	(0.470)	(−0.568)	(−0.047)			

NOTE.—Figures in parentheses are *t*-ratios.

TABLE 3

ANALYSIS-OF-VARIANCE TEST FOR THE CONTRIBUTION OF $P_{t-2}, P_{t-3}, \ldots, P_{t-n}$ TO THE EXPLAINED SUM OF SQUARES

COMMODITY AND PERIOD	df		F-VALUE	SIGNIFICANCE LEVEL
	Numerator	Denominator		
Onions.				
Sept. 1928–Apr. 1942	4	467	10.674	.01
Sept 1948–Oct. 1959	4	363	0.600	*
Nov. 1959–Apr. 1971	4	376	2 702	.05
Potatoes.				
Oct 1925–July 1941	5	611	4.350	.01
Oct 1947–July 1971	5	802	1 463	.25[a]
Pork bellies				
Oct 1949–Sept 1961	4	615	3.050	.05
Oct 1961–Sept 1971	4	511	0 556	*
Hogs				
Oct. 1949–Feb 1966	7	839	2 321	05
Mar 1966–May 1970	7	203	0 456	*
Cattle				
May 1949–Nov. 1964.	8	794	7 887	.01
Dec 1964–July 1971	8	329	2.508	05[b]
FCOJ				
Jan 1957–Oct 1966	9	491	2 918	01
Nov 1966–Aug. 1971	9	231	0 283	*

[a] The F-value for the contribution of $P_{t-3}, P_{t-4}, P_{t-5}, P_{t-6}$ is 1 182, which is not significant at the 25 level

[b] The F-value for the contribution of $P_{t-3}, P_{t-4}, \ldots, P_{t-9}$ is 1 093, which is not significant at the 25 level

* Not significant at 25 level

all traders are not fully informed. For both commodities, P_{t-2} remains highly significant in the periods of futures trading.

Another implication of increased information is a decrease in the variance of the price-forecast error. The empirical counterpart of σ_u^2 in (16) is the standard error of estimate for the regressions in table 2. However, the levels of prices differ between the sample periods, tending to reduce the standard errors in periods of relatively low prices which correspond to periods without futures trading. To remove the price level effect, the standard errors of estimate are expressed as coefficients of variation in table 4. The results support the other evidence of increased information due to futures trading. six of the seven comparisons show a smaller coefficient of variation with futures trading. The conflicting result is for onions, where the coefficient of variation decreases when onion futures are prohibited.

In sum, the tests for each commodity strongly support the hypothesis that additional traders are informed of the latest market information due to futures trading. The evidence on the information effect of futures trading is remarkably consistent over different commodities and time

TABLE 4

STANDARD ERRORS OF ESTIMATE AS PERCENTAGES OF THE SAMPLE MEANS

Commodity and Period	SE of Estimate	Sample Mean	Coefficient of Variation
Onions:			
Sept. 1928–Apr. 1942	0.249	0.804	30.996
Sept. 1948–Oct. 1959	0.358	1.646	21.760
Nov. 1959–Apr. 1971	0.259	1.995	13.006
Potatoes:			
Oct. 1925–July 1941	0.256	1.707	15.014
Oct. 1947–July 1971	0.414	3.323	12.473
Pork bellies:			
Oct. 1949–Sept. 1961	1.728	29.947	5.771
Oct. 1961–Sept. 1971	1.708	31.881	5.357
Hogs:			
Oct. 1949–Feb. 1966	0.667	18.602	3.585
Mar. 1966–May 1970	0.753	22.445	3.355
Cattle:			
May 1949–Nov. 1964	0.704	25.730	2.737
Dec. 1964–July 1971	0.552	27.336	2.019
FCOJ:			
Jan. 1957–Oct. 1966	1.772	39.467	4.490
Nov. 1966–Aug. 1971	1.224	32.950	3.714

periods. This leads me to conclude that a significant price effect of futures trading reflects an increase in market information.

The Returns to a Trading Rule

A major empirical implication from the theory of efficient markets is that a trading rule based solely on the history of a commodity's price will not be profitable if the market's expected price fully reflects current market information. Evidence for the six commodities presented above indicates that those prices do not fully reflect available information in the absence of futures trading. In periods with futures trading, however, prices of four commodities behave as if market information is fully reflected in the price and the other two commodities show increased information. Hence, it is possible that a chartist could earn profits from a trading rule in the periods with no futures trading, but the spot markets become more efficient when there is futures trading and a chartist's profits should be eliminated or sharply reduced. There may not be actual commodity markets where price changes are random—Houthakker (1961) has shown that trading rules for wheat and corn futures are apparently profitable—but the important question for this study is whether there is a difference in profits that is consistent with the evidence on information.

The returns to one trading rule are examined here to see if the predicted change does occur. These returns are gross returns because the calculations do not include some costs such as the cost of transacting. Furthermore, this is a special test because an unlimited number of trading rules could be devised. The trading system considered is the following Estimate an autoregression like those in table 2 for the first half of each sample period and use that equation to forecast the price 1 week hence for each observation in the last half of the sample If the forecast is above the current price, buy one unit of the commodity and sell it at the market price 1 week hence If the forecast is below the current price, sell one unit of the commodity and replace it at the market price 1 week hence The return is measured as the difference between the selling and buying price.

Table 5 contains the statistics pertinent to this analysis First, consider the average returns from the trading system In the absence of futures trading, the average return is positive for all six commodities, but it is negative for three of the commodities when there is futures trading The trading system is risky in every case—none of the mean returns is as much as 25 percent of the standard deviation—however, the variability of returns increases in periods of futures trading for five of the six commodities In order to compare magnitudes, the averages are expressed as annual rates of return on the mean price for each trading period This is approximately the rate of return on investment for a chartist using the trading system With no futures trading, the rates of return range from 11 to 389 percent, with futures trading the range is from −13 to 25 percent The results for onions and potatoes are remarkable Prior to the time of futures trading, the rule yields an annual return of 389 and 108 percent for onions and potatoes, respectively When there is futures trading in these commodities, the returns fall to −1 and 15 percent Notice, too, that when onion futures are prohibited, the return increases substantially to 25 percent Only one of the seven comparisons (FCOJ) shows a greater rate of return with than without futures trading Overall, the difference in returns to the trading rule is consistent with the prediction that spot markets are more efficient because of futures trading. This result is additional support for the hypothesis that traders are more fully informed when there is organized futures trading in a commodity

Could the markets be so imperfect that chartists would not learn of 300, 100, or even 25 percent rates of return, enter the markets, and thereby reduce the returns to more "normal" levels? The answer to this question requires a knowledge of more than the sequence of market prices. The cost of transacting in the spot markets was ignored in calculating the returns; yet costs of identifying potential traders, searching for the best price, and negotiating other terms of the exchange would lower the net returns in every case. Also, it is likely that the cost of transacting per unit of commodity traded declines enough that more than one unit would have

TABLE 5

RETURNS TO A TRADING RULE BASED ON THE HISTORY OF PRICES

Commodity and Period	Average Return per Trade	SD of Returns	Annual Rate of Return	Number of Trades
Onions.				
Sept 1935–Apr 1942	0607	0 261	389 886	240
Mar. 1954–Oct 1959	– 0005	0 368	–1 782	184
Oct. 1965–Apr 1971	0113	0 303	25 610	191
Potatoes				
Apr 1933–July 1941	0297	0 251	108 658	313
Mar 1960–July 1971	0106	0 421	15 652	407
Pork bellies				
Oct 1955–Sept 1961	0942	1 838	17 644	313
Oct 1966–Sept 1971	– 0843	1 863	–13 380	261
Hogs				
Dec 1957–Feb 1966	0385	0 607	11 216	430
Apr 1968–May 1970	0203	0 732	4 494	111
Cattle				
Feb 1957–Nov 1964	0864	0 581	18 188	408
Apr 1968–July 1971	– 0187	0 634	–3 321	174
FCOJ·				
Dec 1961–Oct 1966	1408	1 988	17 986	257
Apr 1969–Aug 1971	1769	1 013	25 889	126

to be traded to make the system worthwhile If so, it is important to know about the breadth of the market, that is, whether an optimum size purchase or sale would change the market price enough to eliminate any potential profits For these reasons, the returns in table 5 do not seem unreasonably high.

III. Conclusion

The empirical evidence on price behavior clearly shows an information effect of futures trading Both week-to-week price analysis and returns to a trading system based on the history of prices yield results consistent with increased information from futures trading So market prices provide more accurate signals for resource allocation when there is futures trading in a commodity Previous work on price effects of futures trading has not investigated the relation between futures trading and market information. Yet the strength and consistency of the evidence reported here suggest that price effects of futures trading result mainly from more fully informed traders.

The results in this study are directly relevant to public policy because they contradict the main argument made in behalf of legislation that prohibits trading in onion futures and regulates futures trading in other commodities The data do not support the claim that price effects of futures trading impose costs on producers, consumers, and others who handle the physical commodity. On the contrary, spot markets seem to work more efficiently because of futures trading. The prohibition of futures trading reduces market efficiency The other restrictions on futures trading may or may not benefit the public—we do not know the effects of this regulation—but the argument for those restrictions is incorrect.

References

Box, G E P., and Pierce, D. A. "Distribution of Residual Autocorrelations in Autoregressive-integrated Moving Average Time Series Models " *J. American Statis. Assoc.* 65 (December 1970) 1509–26

Cox, C. C "The Regulation of Futures Trading " Ph D. dissertation, Univ. Chicago, 1975.

Emery, H. C. *Speculation on the Stock and Produce Exchanges of the United States.* New York· Columbia Univ. Press, 1896.

Fama, E F. "Efficient Capital Markets A Review of Theory and Empirical Work " *J Finance* 25 (May 1970) 383–417.

Gray, R. W. "Onions Revisited." *J Farm Econ.* 65 (May 1963). 273–76.

Hayek, F. A. "The Use of Knowledge in Society." *A.E.R.* 35 (September 1945) 519–30.

Hieronymus, T. A. "Effects of Futures Trading on Prices " In *Futures Trading Seminar* Vol. 1 Madison, Wis Mimir, 1960

Hooker, R. H "The Suspension of the Berlin Produce Exchange and Its Effect upon Corn Prices " *J. Royal Statis Soc* 64 (December 1901) 574–604.

Houthakker, H. S. "Can Speculators Forecast Prices?" *Rev. Econ and Statis.* 39 (May 1957): 143–51.

——— "Systematic and Random Elements in Short-Term Price Movements " *A E R* 51 (May 1961). 164–72

———. "Normal Backwardation " In *Value, Capital, and Growth: Papers in Honour of Sir John Hicks,* edited by James N. Wolfe. Chicago: Aldine, 1968.

Johnson, A C. *Effects of Futures Trading on Price Performance in the Cash Onion Market, 1930–68* U.S Department of Agriculture Technical Bulletin no 1470 Washington Government Printing Office, 1973.

Muth, J F. "Rational Expectations and the Theory of Price Movements." *Econometrica* 29 (July 1961) 315–35

Naik, A. S. *Effects of Futures Trading on Prices.* Bombay: Somaiya, 1970.

Nelson, C R. "Rational Expectations and the Predictive Efficiency of Economic Models." *J. Bus.* 48 (July 1975). 331–43.

Powers, M. J. "Does Futures Trading Reduce Price Fluctuations in the Cash Markets?" *A.E R* 60 (June 1970). 460–64.

Rockwell, C. S. "Normal Backwardation, Forecasting, and the Returns to Commodity Futures Traders." *Food Res. Inst. Studies* 7 (suppl.; 1967): 107–30.

Stewart, B. *An Analysis of Speculative Trading in Grain Futures.* U.S. Department of Agriculture Technical Bulletin no. 1001. Washington: Government Printing Office, 1949.

Taylor, G. S., and Leuthold, R. M. "The Influence of Futures Trading on Cash Cattle Price Variations." *Food Res. Inst. Studies* 13, no. 1 (1974). 29–35.

Telser, L. G. "Futures Trading and the Storage of Cotton and Wheat." *J.P.E.* 66, no. 3 (June 1958): 233–55.

———. "The Supply of Speculative Services in Wheat, Corn, and Soybeans." *Food Res. Inst. Studies* 7 (suppl.; 1967) 131–76

Tomek, W. G. "A Note on Historical Wheat Prices and Futures Trading" *Food Res. Inst. Studies* 10 (1971)· 109–13

Working, H. "Financial Results of Speculative Holding of Wheat" *Wheat Studies Food Res. Inst.* 7 (July 1931): 405–35

——— "Price Effects of Futures Trading" *Food Res Inst Studies* 1, no 1 (1960)· 3–31.

Consequences and Causes of Public Ownership of Urban Transit Facilities

---

B. Peter Pashigian

University of Chicago

The reasons for the shift from private to public ownership of urban transit facilities are the subject of the paper. The regulation theory suggests that this shift is due to the increasing severity of regulation, while the declining-market and externalities hypotheses suggest that increases in automobile ownership are the reason for reduced profits and public ownership Regression results indicate that profit margins of privately owned systems are higher when regulation is by a state rather than a local agency Changes in profit margins over time are found to be directly related to increases in automobile ownership

Introduction

Public production of goods and services it somewhat of an embarrassment to most economists It exists, and will in all likelihood increase in importance, but is difficult to explain An acceptable theory of public production has not yet appeared Several hypotheses have been advanced but have not been pursued or subjected to systematic tests This paper examines the power of these hypotheses to explain the shift from private to public production in the urban transit industry The urban transit industry is a relative newcomer to the area of public production and for that reason is an interesting industry to study It has a long history of predominantly private production along with local or state regulation of fares and routes. In the last decade or so a dramatic shift toward public production has occurred. Among the 117 largest cities in 1970, only 10 had publicly

This study was partially supported by a grant from the Walgreen Foundation and by a grant from the Center for Management of Public and Nonprofit Enterprise, Graduate School of Business, University of Chicago I am indebted to the American Transit Association for supplying the data on transit operations and to the referees for suggesting improvements in the paper Responsibility for errors rests with the author

[*Journal of Political Economy*, 1976, vol 84, no 6]

© 1976 by The University of Chicago. All rights reserved

owned transit systems in 1949. In the next 10 years, only three more systems became publicly owned. From 1960 to 1969, 31 systems were added to the rolls and 15 more became publicly owned from 1970 to 1971. What has changed to cause this shift after so many years of private production? Section I of this paper reviews several hypotheses of the causes of public ownership. Section II describes the sample and the sources and limitations of the data, and Section III introduces the regulation and ownership variables used in the regression analysis The effects of the regulation and ownership variables on profitability are studied in Section IV The comparative performance of public and private transit systems from 1960 to 1970 is examined in Section V. Section VI quantifies the relationship between changes in the size of selected special interest groups and changes in the profit margin of transit firms. The paper concludes with a summary

I. Some Explanations of Public Production

Explanations of public production are few, and those that exist are incomplete This section provides a critical review of some possible candidates

A The Regulation Hypothesis

The regulation hypothesis asserts that regulation is an intervening step before public ownership After a time the regulation of prices and service becomes so stringent that a private firm can no longer earn a normal return Regulators may not permit fare increases to match cost increases or allow unprofitable routes to be discontinued Losses caused by more stringent regulation ultimately result in government ownership. Under government ownership the service is provided with costs exceeding revenues but with public subsidy [1]

The regulation theory lacks an explanation of why regulation becomes more stringent in one city and not in another and, like other theories, assumes but does not explain why net public subsidies for transit will be larger under public ownership than under private ownership. If this were not so, government regulation would be a perfect substitute for public ownership of transit facilities and there would be no need for government ownership.

An extension of the regulation theory requires the identification and measurement of changes in the relative size of groups favoring or opposing more stringent regulation, public ownership, and public subsidies.

---

[1] This view of regulation differs from that of the producer-protection theory, which makes the regulator an agent of the industry (Stigler 1971).

Increases in automobile ownership reduce the demand for urban transit. The reduced demand for urban transit does not, contrary to common opinion, have to result in unprofitable operations and public ownership if the transit firm is allowed to reduce the number of routes and frequency of service and adjust fares. The increase in automobile ownership increases the relative voting strength of nonusers who will oppose public ownership and subsidies and favor profitable transit operations. If workers with low and middle incomes are the primary users of local transit, increases in the size of these groups would be associated with reduced profitability. Firms located in well-developed central business districts (CBD) would also favor low-cost–high-frequency service to the CBD. This group appears to be better organized in the older cities where well-developed businesses and shopping centers exist

The regulation theory assumes that users constitute a voting block large enough to use the political power of the state to redistribute wealth from nonusers in the form of a goods-in-kind subsidy. If users constitute too large a share of voters, any gains from a redistribution would be largely offset by the equivalent higher taxes incurred to cover taxes (Peltzman 1971). It users are too small a fraction of voters, they would have little political strength to affect the policies of the regulator Public ownership of transit facilities did occur first in cities where users represented a larger share of all voters [2] Census data show that the mean proportion of all workers using transit to work in 1960 was 0.28 for systems publicly owned before 1960 and only 0 16 for systems privately owned in 1960 [3] Socialization occurred first in cities where users had greater political strength

[2] If public ownership first occurred in cities where users represented a smaller proportion of all voters, the special interest theory loses its appeal and one would entertain a charity hypothesis The general public offers a goods-in-kind subsidy to a small fraction of the population using local transit

[3] The mean and standard deviation of the proportion of workers using public transit to work in 1960 are shown in the following table

	PROPORTION USING PUBLIC TRANSIT IN 1960	
	Mean	SD
Systems publicly owned before 1960 ($N = 10$)	284	157
Systems privately owned in 1960 ($N = 59$)	163	.083

On the other hand, systems becoming publicly owned since 1959 had a somewhat smaller proportion of workers using public transit in 1960 than systems that remained privately owned from 1960–70. Consequently, the trend to public ownership during the 1960s cannot be simply attributed to the exercise of political strength of users of local transit.

The modern theory of regulation explains the incidence of regulation through an examination of the relative size and relative gains of special interest groups (Stigler 1971). The institutional device used to regulate an industry has in large part been considered a detail. In the transit industry there is some impressive evidence that the political strength of users and nonusers is not independent of the form and type of regulation.[4] The results of a survey of 134 cities are reproduced in table 1 and show that the probability of public ownership is lower when the regulation of fares is by a state public service commission. It appears that state commissions have been less responsive to demands for lower fares or improved service unless state subsidies are provided [5] State governments have been reluctant to provide subsidies to local transit systems in urban centers because of the opposition of voters in rural and smaller urban and suburban communities

If the severity of regulation does vary from city to city and depends in part on the regulating agency, government-owned firms will have smaller profit per vehicle mile (or profit margin) than regulated private firms Under the regulation hypothesis government ownership occurred because regulation did become so stringent that a normal return could no longer be earned.[6] If more stringent regulation occurs when the transit facility is regulated by a local rather than state agency, then profit margins should be higher for firms regulated by state commissions The equality of profit margins for private and public systems constitutes evidence against the regulation hypothesis Higher profit margins for private systems regulated by state rather than local agencies would be consistent with the regulation hypothesis but not predicted by the other hypotheses discussed in this paper

The implications of the regulation hypothesis may be summarized as follows: (1) publicly owned systems should have lower profit margins than privately owned systems, (2) more stringent regulation may be expected when the transit facility is regulated by a local rather than a state agency; and (3) profit margins should decrease less over time, and the probability of a change to public ownership should be lower in cities where the relative size of nonusers, for example, auto owners, has increased.

---

[4] The form and type of regulation may be endogenous and determined by the relative size of the opposing special interest groups However, the type of regulation in urban transit was often determined prior to the period of analysis

[5] Systems regulated by state commissions may have experienced smaller autonomous decreases in the demand for urban transit, experienced smaller decreases in profit margins, and therefore remained under private ownership. Evidence presented below rejects this hypothesis.

[6] Privately owned systems are in markets where regulation is less stringent or where the transition to public ownership has not as yet occurred

TABLE 1

PROBABILITY OF PUBLIC OWNERSHIP AND TYPE OF AGENCY REGULATING FARES

TYPE OF AGENCY REGULATING FARES	TYPE OF OWNERSHIP Private	Public
State commission or agency	38	4
Local agency:		
City council or commission	23	29
Transit authority or other metropolitan authority	3	37
Total	64	70

SOURCE.—R. L. Banks and Associates (1972, p. 88)

B. The Declining-Industry Hypothesis

A more popular hypothesis is that urban transit is a declining industry and, under recent cost conditions, an unprofitable one under any combination of fares and service levels. The declining-industry hypothesis is an incomplete one, for it only explains why the industry is an unprofitable one and not why government ownership must inevitably follow. The demand for many products has decreased without subsequent government ownership of firms. This hypothesis must be combined with another hypothesis, for example, the presence of externalities, if it is to explain why socialization occurs.

The declining-industry hypothesis could be tested by identifying the demand and cost functions in the respective markets and determining if any combination of fares and service levels would yield a normal return. In principle, one could determine whether a normal return could be earned if a system was allowed to abandon less frequently used routes, service the major arteries and streets of a city, and charge higher fares. Such tests are obviously difficult to conduct, so less comprehensive tests must be relied on to evaluate the declining-industry hypothesis. A reduction in demand is more likely to result in losses if there are economies of scale rather than constant cost in the provision of transit services. If there are economies of scale, public ownership should have first appeared in the smaller markets even before transit firms experienced the large loss of riders because of the growth in auto ownership in the post–World War II period. Table 2 shows the ownership status of transit facilities in the 117 most populated cities in 1970. Cities are classified into six groups, and the proportion of cities with public ownership is higher and public ownership occurred earlier in the group with the 20 largest cities. However, cities in this group have higher wage rates, so any cost advantage due to scale may be offset by a higher factor cost. A comparison of groups 3 through 6 shows no systematic variation of the wage rate or in the owned transit systems in 1949. In the next 10 years, only three more systems became publicly owned. From 1960 to 1969, 31 systems were added to the rolls and 15 more became publicly owned from 1970 to 1971. What has changed to cause this shift after so many years of private production? Section I of this paper reviews several hypotheses of the causes of public ownership. Section II describes the sample and the sources and limitations of the data, and Section III introduces the regulation and ownership variables used in the regression analysis. The effects of the regulation and ownership variables on profitability are studied in Section IV. The comparative performance of public and private transit systems from 1960 to 1970 is examined in Section V. Section VI quantifies the relationship between changes in the size of selected special interest groups and changes in the profit margin of transit firms. The paper concludes with a summary

I. Some Explanations of Public Production

Explanations of public production are few, and those that exist are incomplete This section provides a critical review of some possible candidates

A The Regulation Hypothesis

The regulation hypothesis asserts that regulation is an intervening step before public ownership After a time the regulation of prices and service becomes so stringent that a private firm can no longer earn a normal return. Regulators may not permit fare increases to match cost increases or allow unprofitable routes to be discontinued Losses caused by more stringent regulation ultimately result in government ownership Under government ownership the service is provided with costs exceeding revenues but with public subsidy.[1]

The regulation theory lacks an explanation of why regulation becomes more stringent in one city and not in another and, like other theories, assumes but does not explain why net public subsidies for transit will be larger under public ownership than under private ownership. If this were not so, government regulation would be a perfect substitute for public ownership of transit facilities and there would be no need for government ownership.

An extension of the regulation theory requires the identification and measurement of changes in the relative size of groups favoring or opposing more stringent regulation, public ownership, and public subsidies.

---

[1] This view of regulation differs from that of the producer-protection theory, which makes the regulator an agent of the industry (Stigler 1971).

Increases in automobile ownership reduce the demand for urban transit. The reduced demand for urban transit does not, contrary to common opinion, have to result in unprofitable operations and public ownership if the transit firm is allowed to reduce the number of routes and frequency of service and adjust fares. The increase in automobile ownership increases the relative voting strength of nonusers who will oppose public ownership and subsidies and favor profitable transit operations If workers with low and middle incomes are the primary users of local transit, increases in the size of these groups would be associated with reduced profitability. Firms located in well-developed central business districts (CBD) would also favor low-cost–high-frequency service to the CBD. This group appears to be better organized in the older cities where well-developed businesses and shopping centers exist

The regulation theory assumes that users constitute a voting block large enough to use the political power of the state to redistribute wealth from nonusers in the form of a goods-in-kind subsidy If users constitute too large a share of voters, any gains from a redistribution would be largely offset by the equivalent higher taxes incurred to cover taxes (Peltzman 1971). It users are too small a fraction of voters, they would have little political strength to affect the policies of the regulator. Public ownership of transit facilities did occur first in cities where users represented a larger share of all voters [2] Census data show that the mean proportion of all workers using transit to work in 1960 was 0.28 for systems publicly owned before 1960 and only 0.16 for systems privately owned in 1960 [3] Socialization occurred first in cities where users had greater political strength

[2] If public ownership first occurred in cities where users represented a smaller proportion of all voters, the special interest theory loses its appeal and one would entertain a charity hypothesis The general public offers a goods-in-kind subsidy to a small fraction of the population using local transit

[3] The mean and standard deviation of the proportion of workers using public transit to work in 1960 are shown in the following table

	Proportion Using Public Transit in 1960	
	Mean	SD
Systems publicly owned before 1960 ($N = 10$)	.284	.157
Systems privately owned in 1960 ($N = 59$)	.163	.083

On the other hand, systems becoming publicly owned since 1959 had a somewhat smaller proportion of workers using public transit in 1960 than systems that remained privately owned from 1960–70. Consequently, the trend to public ownership during the 1960s cannot be simply attributed to the exercise of political strength of users of local transit.

The modern theory of regulation explains the incidence of regulation through an examination of the relative size and relative gains of special interest groups (Stigler 1971). The institutional device used to regulate an industry has in large part been considered a detail. In the transit industry there is some impressive evidence that the political strength of users and nonusers is not independent of the form and type of regulation.[4] The results of a survey of 134 cities are reproduced in table 1 and show that the probability of public ownership is lower when the regulation of fares is by a state public service commission. It appears that state commissions have been less responsive to demands for lower fares or improved service unless state subsidies are provided[5] State governments have been reluctant to provide subsidies to local transit systems in urban centers because of the opposition of voters in rural and smaller urban and suburban communities

If the severity of regulation does vary from city to city and depends in part on the regulating agency, government-owned firms will have smaller profit per vehicle mile (or profit margin) than regulated private firms. Under the regulation hypothesis government ownership occurred because regulation did become so stringent that a normal return could no longer be earned.[6] If more stringent regulation occurs when the transit facility is regulated by a local rather than state agency, then profit margins should be higher for firms regulated by state commissions The equality of profit margins for private and public systems constitutes evidence against the regulation hypothesis Higher profit margins for private systems regulated by state rather than local agencies would be consistent with the regulation hypothesis but not predicted by the other hypotheses discussed in this paper

The implications of the regulation hypothesis may be summarized as follows· (1) publicly owned systems should have lower profit margins than privately owned systems; (2) more stringent regulation may be expected when the transit facility is regulated by a local rather than a state agency, and (3) profit margins should decrease less over time, and the probability of a change to public ownership should be lower in cities where the relative size of nonusers, for example, auto owners, has increased

---

[4] The form and type of regulation may be endogenous and determined by the relative size of the opposing special interest groups However, the type of regulation in urban transit was often determined prior to the period of analysis

[5] Systems regulated by state commissions may have experienced smaller autonomous decreases in the demand for urban transit, experienced smaller decreases in profit margins, and therefore remained under private ownership. Evidence presented below rejects this hypothesis.

[6] Privately owned systems are in markets where regulation is less stringent or where the transition to public ownership has not as yet occurred.

TABLE 1

PROBABILITY OF PUBLIC OWNERSHIP AND TYPE OF AGENCY REGULATING FARES

TYPE OF AGENCY REGULATING FARES	TYPE OF OWNERSHIP	
	Private	Public
State commission or agency	38	4
Local agency.		
City council or commission	23	29
Transit authority or other metropolitan authority	3	37
Total	64	70

SOURCE.—R. L. Banks and Associates (1972, p. 88)

B. *The Declining-Industry Hypothesis*

A more popular hypothesis is that urban transit is a declining industry and, under recent cost conditions, an unprofitable one under any combination of fares and service levels The declining-industry hypothesis is an incomplete one, for it only explains why the industry is an unprofitable one and not why government ownership must inevitably follow The demand for many products has decreased without subsequent government ownership of firms This hypothesis must be combined with another hypothesis, for example, the presence of externalities, if it is to explain why socialization occurs

The declining-industry hypothesis could be tested by identifying the demand and cost functions in the respective markets and determining if any combination of fares and service levels would yield a normal return In principle, one could determine whether a normal return could be earned if a system was allowed to abandon less frequently used routes, service the major arteries and streets of a city, and charge higher fares Such tests are obviously difficult to conduct, so less comprehensive tests must be relied on to evaluate the declining-industry hypothesis A reduction in demand is more likely to result in losses if there are economies of scale rather than constant cost in the provision of transit services. If there are economies of scale, public ownership should have first appeared in the smaller markets even before transit firms experienced the large loss of riders because of the growth in auto ownership in the post–World War II period Table 2 shows the ownership status of transit facilities in the 117 most populated cities in 1970 Cities are classified into six groups, and the proportion of cities with public ownership is higher and public ownership occurred earlier in the group with the 20 largest cities However, cities in this group have higher wage rates, so any cost advantage due to scale may be offset by a higher factor cost. A comparison of groups 3 through 6 shows no systematic variation of the wage rate or in the

TABLE 2

OWNERSHIP, FARE, WAGE RATE, AND FORM OF REGULATION BY SIZE OF CITY

GROUP	PERCENT PUBLIC OWNERSHIP	MEDIAN DATE OF PUBLIC OWNERSHIP	MEAN HOURLY WAGE RATE*	FORM OF REGULATION (WHEN KNOWN)		
				State	Authority	City Council
1. Largest 20	70	1958	$3.88(18)	4	11	4
2. Second largest 20	50	1968	3 59(19)	4	8	5
3. Third largest 20	55	1967	2.89(17)	4	5	9
4. Fourth largest 20	55	1967	3 24(16)	4	3	9
5. Fifth largest 20	30	1970	2 99(15)	8	0	6
6. Next 17	41	1967	2.99(10)	7	3	2

NOTE.—Figures in parentheses denote sample size.
* Wage rate is union wage rate for operator of one-man bus with 1 year of experience in 1970

proportion of cities with public ownership of transit facilities with the size of the city. Some further confirming evidence is the absence of any economies of scale in a cross-sectional analysis of the operating costs of 58 systems.[7]

One other test clearly distinguishes between the declining-industry hypothesis and the regulation hypothesis. The declining-industry hypothesis predicts that a rise in auto ownership will lower the profitability of transit firms and increase the probability of change to public ownership, while the regulation hypothesis makes the opposite prediction. Hence, the observed relation between the change in profit margin over time and change in automobile ownership provides evidence to distinguish between the competing hypotheses.

C. The Externalities Hypothesis

The trip cost of a transit passenger includes a money and time component. Even though bus services are produced under constant returns to scale, trip cost per ride will decrease with an increase in the number of passengers if the time between bus arrivals (waiting time) decreases as the number of passengers and buses increases (Mohring 1972) If so, a permanent decrease in the number of riders imposes an external diseconomy on the remaining riders by increasing waiting time.[8] If long-run average cost per trip decreases with the number of passengers (where cost includes time and money costs), marginal cost per trip will be less than average cost per trip. In the absence of any subsidy, the equilibrium number of trips will be determined where the average cost per trip equals the demand price per trip if a uniform fare is charged. If the demand price is to be equated to the marginal cost of a trip, a subsidy is required and may take the form of reducing the fare (the money cost of a trip) below the bus cost per trip. By reducing the fare below bus cost per trip, it may be possible to equate the demand price per trip to the marginal cost (bus plus time cost) of an additional trip. Public ownership could be justified as a method of practicing marginal cost pricing. Why a private transit firm could not be subsidized is unclear

---

[7] The OLS regression results for 1970 were $\log C = -1.70(8.2) + 1.16(44.4) \log VM + 12(1\ 8) \log WAGE - 34 \times 10^{-7}(8)\ (VM \log VM) + u$, where C denotes operating expenses and −.07 denotes 0000000, VM denotes miles, and WAGE denotes hourly wage of union bus hires with 1 year of experience. The cross-product term determines if the returns-to-scale parameter depends on total vehicle miles supplied. Numbers in parentheses are *t*-ratios The results fail to reveal any significant economies of scale. Two-stage results yielded similar results

[8] Mohring (1972) has presented an extensive discussion of this argument. He shows that waiting time per trip will decrease as the number of passengers and number of buses increase proportionally if the average wait for service is proportional to the time between bus arrivals.

The externalities hypothesis is an incomplete explanation of socialization of transit firms because it does not explain which cities will offer subsidies The argument is so general that all systems should be offered subsidies if the goal is to achieve marginal cost pricing To give content to the hypothesis and explain the shift to public ownership, it is necessary to identify systems where the discrepancy between marginal and average cost per trip may have increased most during the last decade or so. Mohring (1972) suggests that this discrepancy increases with decreases in the number of passengers per mile hour (at least over some range) If the intensity of use of transit facilities decreases with increases in auto ownership or decreases in bus vehicle miles supplied, then cities with larger increases in auto ownership per household or with larger percentage decreases in bus vehicle miles may be expected to have a higher probability of a change to public ownership In summary, the externalities hypothesis predicts that (1) the profit margin (before subsidy) should be lower for publicly owned than for privately owned systems,[9] and (2) increases in auto ownership should be inversely related to changes in the profit margin and directly related to the probability of a change to public ownership The predictions of the externalities hypothesis are similar to those of the declining-industry hypothesis and therefore differ from the regulation hypothesis

II. Selection of Sample

As a starting point, the 117 largest cities in 1970 were selected for study The ownership status of the transit system in each city at the end of 1971 was determined along with the date of transition to public ownership when applicable The data base was augmented by including economic and demographic data from current and past issues of the *Census of Population* and other sources Data describing the regulatory agencies were obtained from the study by R L Banks and Associates (1972) The definition of variables and sources of data are presented in the Appendix

Revenue and cost data of transit facilities were supplied by the American Transit Association (ATA). In many cases, the ATA data are incomplete because not all systems provide continuous reports and some report unexpected and unexplained year-to-year differences These limitations of the ATA data reduced the sample so that 58 systems were used in the 1970 cross sections. Obviously, the sample was not selected by random sampling While this is a major limitation, the sample is the largest that may be studied with available information. A second sample included 40 of the above 58 systems that reported data in both 1960 and 1970 so that comparisons over time could be made. Not all systems reporting to the ATA use the same accounting conventions or systems of accounts. One limitation of the study is that revenue and cost measures may not be comparable across transit systems Some tests are formulated in terms of changes over time to reduce the effects of differences in accounting methods between transit systems[10]

---

[9] The profit margin should decline after the system has become publicly owned and a public subsidy is provided.

III. Classification by Regulation and Ownership

Each transit system was classified into one of seven groups which identified regulatory agency and ownership form Systems are regulated by three general forms (1) state commission (ST, usually a public service commission), (2) local government (CC, often a city council), and (3) a transit authority or district (AU) Publicly owned systems were also classified into two subclasses (*a*) public ownership occurred before 1960 or (*b*) after 1959[11] The notation used is summarized below

CC = regulated by city council or other form of local government,
PCCA = regulated by city council and publicly owned since 1959,
PCCB = regulated by city council and publicly owned before 1960,
ST = regulated by state commission,
PSTA = regulated by state commission and publicly owned since 1959,
PAUB = regulated by transit authority or district and publicly owned before 1960,
PAUA = regulated by transit authority or district and publicly owned since 1959

Except for PAUA, each of these variables is introduced as a dummy variable The PAUA is treated as the base variable Therefore, the coefficients of the other seven dummy variables represent the effects of regulation and ownership relative to the effects of PAUA The effect of regulation without public ownership is measured by the coefficients of CC or ST For example, if private regulated firms receive greater revenue

[10] Operating revenues and operating costs measure revenues and costs from operations only Nonoperating revenues are excluded Operating costs include maintenance, depreciation, taxes, and other standard cost items This cost measure undoubtedly contains an arbitrary and random component since the depreciation figures will depend on the age and method of depreciation of equipment However, measurement error in the cost figures may not be serious Further analysis showed that the change in profit margin from 1960 to 1970 was an important determinant of the change in the probability of public ownership If there were serious distortions in the revenue and cost figures, such a relationship would not have been observed

[11] Three exceptions were made A private transit firm was regulated by a transit authority or district in Buffalo, Nashville, and Washington, D C This is considered a temporary arrangement in Buffalo and Nashville since public ownership was anticipated in the near future. The system in Washington, D C, appeared to have many characteristics of a publicly owned system. Consequently, these three systems were included in the PAUP class even though a privately owned firm was ostensibly operating

per vehicle mile than do systems in PAUA (publicly owned since 1959 and regulated by a transit authority or district), the coefficients of ST and CC will be positive. If regulation by state commission is less stringent than by city council, the coefficient of ST will be larger than the coefficient of CC. Given the form of regulation, the effect of public ownership is measured by the coefficients of PCCA and PCCB for systems regulated by a city council and by PSTA for systems regulated by state commissions. The variables are defined so that the combined effects of regulation and public ownership are measured by the sum of the coefficients of CC and PCCA (or PCCB) and by the sum of the coefficients of ST and PSTA[12] If the effects of public ownership are independent of the regulatory agency, the sum of these coefficients should not differ from zero The coefficient of PAUB measures the effect of public ownership for systems under public ownership before 1960 and regulated by transit authority or district. If the stringency of regulation has increased with the number of years under public ownership, the coefficient of PAUB will be negative The different tests for measuring the effects of regulation and ownership are summarized in table 3.

IV. Effect of Regulation and Ownership on Profitability

The initial regression results show how the regulation and ownership variables affect (*a*) revenue per vehicle mile in 1970, RVM, given operating expense per vehicle mile in 1970, CVM, and (*b*) the profit margin, PM, given vehicle miles supplied in 1970, VM The regression equations are of the form

$$\begin{aligned} \text{RVM} &= a_0 + a_1\text{CVM} + \sum_{i=1}^{6} a_i D_i + u \\ \ln \text{RVM} &= b_0 + b_1 \ln \text{CVM} + \sum_{i=1}^{6} b_i D_i + v \qquad (1) \\ \text{PM} &= c_0 + c_1 \ln \text{VM} + \sum_{i=1}^{6} c_i D_i + w, \end{aligned}$$

where D_i denotes the *i*th dummy variable and ln denotes the logarithm of a variable

Under the regulation hypothesis the coefficients of ST and CC should be positive and significantly different from zero if regulation has been less stringent for privately owned firms. The coefficient of ST should exceed that of CC if regulation is less stringent under state commission The coefficients of PCCA and PTSA should be negative and significantly different from zero. Finally, the coefficients of PAUB and PCCB may be

---

[12] The class, PTSB, was empty.

TABLE 3

MEASURES OF THE EFFECTS OF REGULATION AND OWNERSHIP

EFFECT	MEASURED BY	
	Coefficient of	Sum of Coefficients of
Regulation by state government (private ownership)	ST	
Regulation by state government and public ownership after 1959		ST and PSTA
Regulation by local government (private ownership)	CC	
Regulation by local government and public ownership before 1960		CC and PCCB
Regulation by local government and public ownership after 1959		CC and PCCA
Regulation by transit authority with public ownership before 1960	PAUB	

zero or negative depending on the effect of years of public ownership on RVM

The results in columns 1 and 2 of table 4 show that the coefficient of ST is positive and statistically significant at the 2 5 percent level.[13] The coefficient of CC is positive, but it is not statistically significant at the 5 percent level The coefficients of PCCA, PCCB, and PSTA are all negative and significant at the 5 percent level in column 2 but not in column 1 Systems regulated by transit authorities before 1960 received significantly lower RVM than systems regulated by transit authorities after 1959. The dependent variable in column 3 is the operating profit margin, operating revenue less operating cost divided by operating revenue Similar qualitative conclusions may be drawn of the effects of the regulation and ownership variables [14]

Knowing the type of regulating agency and the form of ownership is valuable information in predicting the financial position of a transit system. Privately owned systems regulated by state commissions receive higher RVM and earn higher operating margins than publicly owned systems regulated by state commissions Regulation by state commission appears less restrictive than by local government. Privately owned systems regulated by local government did not receive significantly higher RVM or earn higher profit margins. Systems publicly owned before 1960 have lower RVM and profit margins Systems publicly owned after 1959 tend to have lower RVM and profit margins than privately

---

[13] For a one- (two-) tailed test, a t-value of 1 67 (2 00) is significant at the 5 percent level with 60 observations

[14] The direct relationship between vehicle miles and profit margin is surprising. Larger systems would be expected to utilize their equipment more efficiently and obtain higher revenues per dollar of assets and therefore earn a lower profit margin if a normal rate of return is earned. The evidence suggests that larger systems are utilizing their equipment less efficiently than smaller systems

TABLE 4

REVENUE PER VEHICLE MILE AND OPERATING PROFIT MARGIN

Variable	Dependent Variable RVM, 1970 (1)	ln (RVM, 1970) (2)	Operating Profit Margin (3)
Constant	0.17	−0.15	−1.09
	(2.2)	(3.0)	(4.6)
Operating expense per VM, 1970 (CVM or ln [CVM])	0.67	1.02	
	(8.6)	(10.2)	
CC (N = 21)*	0.05	0.11	0.11
	(0.8)	(1.3)	(1.2)
ST (N = 15)	0.11	0.18	0.23
	(2.0)	(2.5)	(2.9)
PCCA (N = 10)	−0.10	−0.18	−0.12
	(1.4)	(1.9)	(1.1)
PSTA (N = 3)	−0.12	−0.24	−0.23
	(1.3)	(2.0)	(1.7)
PAUB (N = 6)	−0.11	−0.18	−0.45
	(1.4)	(1.9)	(4.0)
PCCB (N = 4)	−0.13	−0.23	−0.31
	(1.4)	(2.0)	(2.4)
Log of vehicle miles, 1970			0.10
			(3.9)
R^2	.668	.734	.461
σ_u	.147	.183	.208
N	58	58	58

NOTE.—Figures in parentheses denote t-ratios.

* For illustration, there were 21 systems regulated by local government of which 10 became publicly owned since 1959 and four were publicly owned before 1960. Thus, seven systems were privately owned and regulated by local government at the end of 1971. Incidentally, the base class PAUA contains 13 observations.

owned firms, but the precision of some of the coefficient estimates is low. The evidence does not contradict the regulation hypothesis and in some ways is consistent with it. While the declining-industry and externalities hypotheses are also capable of explaining why the profit margins of publicly owned firms are lower than privately owned firms, they do not explain why privately owned firms under state commission earn significantly higher margins.

A trend toward more stringent regulation can be detected by comparing the coefficients of the dummy variables in successive cross sections. If regulation has become more stringent over time, systems that become publicly owned after 1960 (and possibly for systems which had already become publicly owned) should have experienced larger decreases in RVM relative to CVM from 1960 to 1970. Separate cross-section results are reported in columns 1 and 2 of table 5 for the 40 systems that reported operating data in both years. There is no significant difference between the systems in 1960 except for those systems publicly owned before 1960 and regulated by a transit authority. Hence, systems that were to become

TABLE 5

REVENUE PER MILE AND OTHER OPERATING CHARACTERISTICS OVER TIME

			CHANGE FROM 1960 TO 1970 IN		
VARIABLE	l RVM, 1960 (1)	l RVM, 1970 (2)	Profit Margins (3)	Log of Operating Cost (4)	Log of Revenue per Passenger (5)
Constant	0.01	−0.11	−0.20	0.33	0.48
	(0.2)	(2.1)	(3.9)	(8.8)	(7.8)
Log CVM	0.87	0.96			
	(12.5)	(8.3)			
CC ($N = 15$)	−0.05	0.11	0.13	−0.12	0.00
	(1.0)	(1.1)	(1.4)	(1.8)	(0.0)
ST ($N = 10$)	0.00	0.13	0.11	−0.01	0.07
	(0.1)	(1.7)	(1.4)	(0.1)	(0.8)
PCCA ($N = 7$)	−0.01	−0.25	−0.31	0.10	−0.24
	(0.1)	(2.6)	(3.2)	(1.5)	(2.2)
PSTA ($N = 1$)	0.04	0.05	0.00	−0.14	−0.35
	(0.4)	(0.3)	(0.0)	(1.0)	(1.8)
PAUB ($N = 5$)	−0.13	−0.24	−0.14	0.17	0.04
	(2.7)	(2.4)	(1.5)	(2.6)	(0.3)
PCCB ($N = 3$)	0.05	−0.18	−0.22	0.17	−0.17
	(0.8)	(1.5)	(1.9)	(2.0)	(1.2)
Change in log of vehicle miles			−0.22	0.88	−0.09
			(1.8)	(10.1)	(0.6)
R^2	.877	.767	.478	.858	.328
σ_u	.083	.166	.162	.117	.189
N	40	40	40	40	40

NOTE.—Figures in parentheses denote *t*-ratios.

publicly owned sometime after 1959 received the same RVM in 1960 (given CVM) as systems that subsequently remained privately owned after 1959. Ten years later, some changes have occurred. Relative to the base group (PAUA), privately owned systems under regulation by state commissions received significantly higher revenues per vehicle mile.[15] Publicly owned systems (except for the one observation in PSTA) received a lower RVM (relative to CVM) and had suffered a relative deterioration in revenue received per vehicle mile. While the significance level of the estimated coefficients of CC, PSTA, and PCCB in the 1970 regression is below the 5 percent level, the direction of change in the value of the coefficient of each dummy variable from 1960 to 1970 suggests a relative improvement in RVM for privately owned systems and a relative deterioration for publicly owned systems. This relative deterioration is not due to, say, a larger autonomous decrease in the demand for transit services in just those cities with systems that became publicly owned after 1959 because the deterioration was also experienced by systems that were publicly owned before 1960. While these results are not inconsistent with the regulation hypothesis, they are weaker because of the reduced significance levels of some of the coefficients.

---

[15] The coefficient of ST is significant at the 5 percent level given a one-tailed test.

In columns 3–5, changes between 1960 and 1970 in (1) the profit margin, (2) the log of operating cost, and (3) the log of revenue per passenger are related to the ownership and regulation dummies. The change in the profit margin shows an improvement for the privately owned firms relative to the systems that become publicly owned after 1959 and relative to the systems that were publicly owned before 1960. But, once again, the significance level of the coefficients of the ST and PAUB dummy variables falls below the 5 percent level. The coefficients of only PCCA and PCCB are negative and significant at the conventional 5 percent level of significance The decline in the profit margin for publicly owned systems cannot be attributed to any one cause Operating cost grew most rapidly for systems in the PAUB class, while revenue per passenger grew less rapidly for systems in the PCCA and PSTA classes.

If public systems raise fares and reduce service by a smaller percentage than private systems do, they would have experienced smaller percentage declines in the number of revenue passengers and in vehicle miles supplied Cities with public transit systems should experience a smaller percentage increase in automobiles since the money and time price of local transit would have risen less rapidly over time. These hypotheses were tested. The results (not presented) indicated no significant difference between public and private systems in the percentage change from 1960 to 1970 in revenue passengers or in vehicle miles supplied. The growth in autos per household was significantly higher in those cities where transit systems remained privately owned from 1960 to 1970 than in cities where systems became publicly owned. The percentage increase in auto ownership was largest, after controlling for the growth of households, in cities with private systems regulated by state commissions. Publicly owned systems performed less well from 1960 to 1970, but not because they were in markets that experienced unusually large decreases in the demand for urban transit. To review, privately owned systems appeared to have higher revenue per vehicle mile and profit margins in 1970 than publicly owned systems. While the relative profit performance of privately owned systems appears to have improved over time, the precision of these results is clearly lower. The larger growth rate of auto ownership in cities with privately owned systems is consistent with the regulation hypothesis The evidence also indicates the profit margins are higher when regulation is by state commission rather than a local regulatory agency. The absence of a larger percentage decrease in revenue passengers and vehicle miles for privately owned systems constitutes opposing evidence.

V. Comparative Performance of Public and Private Transit Systems

The comparative performance over time of private and public transit systems will reveal whether publicly owned systems have been more successful than privately owned systems in reducing the diversion of passengers from local transit to the private auto A comparative analysis of this type also provides additional tests of the competing hypotheses for public ownership. If publicly owned systems charge lower prices or provide more frequent service and require less waiting time, the number of users should have declined less under public than under private ownership.

The performance of a transit system cannot be easily summarized by any one output measure Nevertheless, policy discussions are often directed at the question of the relative use of local transit The performance measure chosen for study in this section is the census estimate of the percentage of city workers using public transit to work. While this measure is not without its defects, it is central to current policy issues in urban transit [16]

Let p_t be the percentage of city workers who use public transit to go to work. This percentage is determined by a vector of independent variables x_t through

$$p_t = \frac{100}{1 + e^{f(x_t) + u_t}}, \tag{2}$$

where u_t denotes a disturbance term The exponent of e may extend from minus infinity to plus infinity, so that p_t is bounded by zero and 100. A transformation of (2) gives

$$y_t \equiv \ln\left[\frac{p_t}{100 - p_t}\right] = -[f(x_t) + u_t] \tag{3}$$

The form of f is given by

$$f(x) = \alpha_0 + \alpha_1 L_t + \alpha_2 M_t + \alpha_3 H_t + \alpha_4 AH_t + \alpha_5 WH_t + \alpha_6 lHO_t + \cdot\ \cdot, \tag{3a}$$

where $\alpha_1 < 0$, $\alpha_2 = ?$, $\alpha_3 > 0$, $\alpha_4 > 0$, and $\alpha_5 < 0$, L_t denotes the percentage of families in the city with low family incomes (defined in Appendix); M_t denotes the percentage of families with middle incomes;

[16] A public transit system may increase ridership by providing frequent and subsidized bus service, while a private firm suffers a reduction in ridership because its service is not subsidized. The improved performance of the public system would then be due to the subsidy and not necessarily to public ownership. Other limitations of this measure are (1) the exclusion of users other than workers, and (2) differences between cities in the distribution of jobs in and outside of the central business district.

H_t denotes the percentage with high incomes; AH_t denotes autos per household; WH_t denotes workers per household; and lHO_t denotes the log of the number of households.[17] Since $L_t + M_t + H_t = 100$, equation (2) may be rewritten as

$$f(x) = \alpha_0 + \alpha_2 \cdot 100 + (\alpha_1 - \alpha_2)L_t + (\alpha_3 - \alpha_2)H_t + \alpha_4 AH_t + \cdots \tag{4}$$

The coefficient of L_t or H_t measures the net effect of a 1 percent change in the percentage of low- or high-income families on y_t. The regulation and ownership variables are also included in (3a) to test for the effects of ownership and regulation on the percentage using public transit

Regression results for equation (3) are presented in table 6 [18] The percentage of workers using local transit decreases as the number of autos per household rises and increases as the number of households in the urbanized area rises The number of workers per household increases transit usage in both regressions, as expected, although the estimated coefficient in the 1970 regression is only statistically significant at the 10 percent level A 1 percent point increase in L_t (percent of families with low income), holding the percentage with high income constant, increases the percentage of workers using public transit in the 1970 cross section but has no significant effect in 1960 [19]

The regulation and ownership variables are not significant determinants of the percentage of workers using local transit The market share of local transit in cities with systems becoming publicly owned after 1959 and regulated by CC is significantly lower in 1970 The results for both cross sections do not show a larger percentage of workers using local transit under public ownership. A comparison of the estimated coefficients for 1960 and 1970 for each dummy variable leaves the impression that privately owned systems outperformed the publicly owned systems from 1960 to 1970 (the exception is the PSTA class), even though the coefficients of the individual dummy variables are not significant [20] The data do reject the hypothesis that publicly owned systems improved their performance from 1960 to 1970 relative to privately owned systems

---

[17] AH_t, WH_t, and lHO_t are measured over the urbanized area since city data were not readily available

[18] Because transit firm operating data were not used in this analysis, a larger sample size was selected and used

[19] The absence of an effect in the 1960 cross section may be due to the use of the poverty measure as low income in 1970 while $3,000 is the upper bound for the low-income class in 1960.

[20] The coefficient of PAUB and PCCB decreased from 1960 to 1970, which indicates a deterioration in the performance of systems under public ownership before 1960. The deterioration in performance is not limited to systems that only recently came under public ownership. Further tests rejected the hypothesis that the performance of publicly owned systems increases with the number of years under public ownership.

TABLE 6

DETERMINANTS OF THE PERCENTAGE OF WORKERS USING PUBLIC TRANSIT—LOGIT MODEL

VARIABLE	1960		1970	
	Coefficient	*t*	Coefficient	*t*
Constant	−2 54	2 6	−0 71	0 6
Percent with low income	0 001	0 1	0 04	2 0
Percent with high income	−0 02	1 0	0.02	1 2
Autos per household	−2 88	7 6	−4 63	10.9
Workers per household	1.67	3 2	0 99	1.4
Log of number of households	0 39	7 4	0 42	6.0
CC (N = 24)	0 18	1 4	0 22	1.3
ST (N = 19)	−0 02	0 2	0.09	0.6
PCCA (N = 12)	−0 05	0 3	−0 36	2 0
PSTA (N = 3)	0 04	0 2	0 13	0 5
PAUB (N = 7)	0 15	0 8	−0 12	0 6
PCCB (N = 3)	−0 06	0 3	−0 13	0.5
R^2	822		845	
σ_u	328		396	
N	70		70	

If the goal of public ownership is to increase ridership by improving service frequency and, indirectly, reducing trip cost by reducing waiting time, the goal was not realized from 1960 to 1970. If public ownership did not involve significant public subsidies, the apparent failure of publicly owned transit systems to improve relative performance may not be surprising. Subsidization of capital purchases and of operating cost appears to have become more popular in the second half of the decade. Perhaps the full effects of public subsidies have yet to be experienced, and future results may yet reveal an improvement in performance under public ownership The results of the comparative-performance analysis are also inconsistent with the regulation hypothesis The regulation hypothesis predicts a larger percentage decline in transit users under private rather than public ownership These and the earlier tests fail to reveal such an effect[21]

VI. Determinants of the Probability of a Change to Public Ownership

As reported above, the operating profit margin of many, but not all, systems declined from 1960 to 1970 If the decline in margin was ultimately caused by the increased political strength of groups favoring public ownership, it should be possible to show a relationship between the change in margin and the change in the size of each special interest group

---

[21] If public ownership occurred because the performance of the transit firm had deteriorated, the coefficients of PCCA and PSTA would have declined relative to CC and ST from 1960 to 1970 and the coefficients of CC and ST should have increased relative to the base group.

Several special interest groups will exert political support for, or opposition to, more stringent regulation, public ownership, and public subsidies to mass transportation. Increases in the number of autos per household will increase the number of nonusers, a majority of which favor profitable operations and oppose subsidies and public ownership. Increases in the number of workers per household increase the number of users, a majority of which will favor lower fares and subsidized service The effect of increases in the percent of upper-income families in a city is ambiguous. If upper-income families are nonusers, they will favor profitable operations and oppose higher taxes and subsidies to mass transit. If upper-income families are heavy users, they may favor more frequent and subsidized service since the higher the opportunity cost, the higher the family income. Firms located in the central business district would favor low fares and frequent service to the CBD. Given the size of the city, cities with lower growth rates are typically older and have larger well-defined CBDs Systems in cities with lower growth rates are expected to have larger declines in profit margins Finally, the size of city is introduced to determine if the cost of organizing political activities is higher in larger cities because of the free rider problem.

The change in the profit margin from 1960 to 1970 for the systems in the 40-city sample was regressed on the change in the size of the special interest groups described above. The results are shown in table 7. After controlling for changes in vehicle miles, smaller decreases in profit margin were experienced by transit systems located in urban areas with (*a*) larger increases in autos per household, (*b*) larger increases in the percent of upper-income families, (*c*) smaller increases in workers per household, and (*d*) smaller household growth rates. The long-term population growth rate from 1950 to 1970 and the size of urban area have no significant effect on the change in margin [22]

These results are not easily reconciled with the declining-market or externalities hypothesis If the fall in profit margin is caused by the decrease in the demand schedule for transit services, an increase in autos per household, a decrease in workers per household, or a decrease in the number of households should have been associated with a decrease in the margin. In this sample of 40 cities, increases in autos per household, decreases in workers per household, and decreases in the number of households are associated with increases in the profit margin These results are consistent with the hypothesis that nonusers support profitable

---

[22] Additional regression results showed that the change in autos per household was inversely related to the change in the probability of public ownership. Hence, cities with larger increases in autos per household had a lower probability of changing from private to public ownership of transit facilities between 1960 and 1970. For further results, see Pashigian (1975).

TABLE 7

CHANGE IN PROFIT MARGIN, 1960–70

Variable	Est.	t	Est	t	Est.	t
Constant	−0.64	3.2	−0.33	3.8	−0.63	2.9
Δ in autos per household	1.37	3.3	0.67	2.5	1.47	3.7
Δ in workers per household	−0.43	1.7	−0.61	2.4		
Δ in percent with upper income . .	0.014	2.5	0.013	2.2	0.016	2.6
Δ in log of vehicle miles .	−0.28	2.6	−0.32	3.0	−0.23	2.1
Δ in log of households, 1960–70 .	−0.24	1.7			−0.28	1.9
Ratio of city population, 1950–70			−0.039	0.9		
Δ in percent of population between 18–64 years					−0.026	1.2
Log of households, 1960	0.031	1.1			0.016	0.5
R^2 .	.535		.471		.517	
σ_y	150		158		153	
N .	40		40		40	.

NOTE.—Δ = difference.

transit operations and oppose policies which will result in higher taxes and public subsidies.[23] The evidence suggests that automobile owners and high-income groups oppose socialization of transit facilities. The increase in automobile ownership has been the major reason for the decline of urban transit but, at least in recent times, has also been an important factor in maintaining profit margins of transit firms and limiting government ownership of transit facilities.

Summary

The results of this paper suggest that the transition from a private regulated transit firm to a government-owned firm is not a mere cosmetic change as far as the financial position of the transit system is involved. In general, government ownership is associated with lower profit margins and lower revenue per vehicle mile. Government ownership of transit facilities yields results which were unattainable under government regulation and explains why there is a demand for government ownership. The deterioration of profit margins and the increase in the number of government-owned systems strongly suggest that private transit firms did not benefit from government regulation during the last 15 to 20 years. With some exceptions, the regulatory agencies have been captured not by the transit firms of the industry but by the riders

---

[23] It is possible that a decline in ridership reduces capacity utilization (by lowering revenue per dollar of capital invested) and causes the profit margin to rise so that the rate of return on capital does not fall. An increase in autos per household may be directly related to the change in profit margin because of lower utilization of equipment when ridership declines. If the change in margin solely reflected the change in utilization, the change in margin would not be an important determinant of the change in the probability of public ownership (see n. 10 above).

While a final judgment may be premature, the experience of the 1960s should dampen hopes that public ownership of transit facilities will moderate the diversion of transit users to the automobile Follow-up study is needed, but, for the moment, the results indicate that government-owned facilities have not outperformed privately owned facilities in maintaining ridership

While the decline in the profit margin of transit systems from 1960 to 1970 was pervasive, it was smaller or negligible in urban areas where the political strength of nonusers increased most Transit systems with smaller declines in profit margins were located in urban areas with larger increases in autos per household and in the percent of families with upper incomes and smaller increases in workers per household The evidence appears to contradict the common view that increases in automobile ownership are responsible for unprofitable transit operations The rise of the automobile is the major cause of the decline of the urban transit industry, but the results of this paper suggest that it is not responsible for unprofitable transit operations and appears to be an important factor inhibiting the expansion of government ownership of transit facilities

Appendix

Definitions of Variables and Sources of Data

Variable	Source
A Ownership and regulatory variables	
1. CC = regulated by city council or other form of local government,	American Transit Association and R L. Banks and Associates (1972)
2 PCCA = regulated by city council and publicly owned since 1959,	
3 PCCB = regulated by city council and publicly owned before 1960,	
4 ST = regulated by state commission,	
5 STA = regulated by state commission and publicly owned since 1959,	
6. PAUA = regulated by transit authority or district and publicly owned since 1959,	
7 PAUB = regulated by transit authority or district and publicly owned before 1960	
B. Transit system variables	
1. RVM = operating revenue per vehicle mile,	American Transit Association (computer printouts)
2. CVM = operating cost per vehicle mile,	
3. PM = operating profit margin,	
4. VM = vehicle miles.	

C. Economic and demographic variables

1. Percentage of families with low income a) Income less than $3,000 in 1960, b) Income less than poverty income in 1970	U S Bureau of the Census (various)
2. Population of city, 1950, 1960, and 1970	
3 Percentage of families with high income a) Income greater than $10,000 in 1960 b) Income greater than $15,000 in 1970	
4 Autos per household	Department of Transportation (1971, 1973)
5 Workers per household.	
6 Number of households	
7 Percent of workers using public transit	

References

Banks, R. L , and Associates *Study and Evaluation of Urban Mass Transportation Regulation and Regulatory Bodies* Springfield, Va Nat Tech Information Service, 1972

Mohring, Herbert "Optimization and Scale Economies in Urban Bus Transportation " *A E R* 62 (September 1972) 591–604

Pashigian, B Peter "Consequences and Causes of Public Ownership of Urban Transit Facilities " Report no 7536, Center Math Studies Bus and Econ Univ. Chicago, 1975

Peltzman, Sam. "Pricing in Public and Private Enterprises Electric Utilities in the United States " *J Law and Econ.* 14 (April 1971) 109–65

Stigler, George "Theory of Regulation " *Bell J Econ and Management Sci* 2 (Spring 1971) 3–21

U S. Bureau of the Census *U S Census of Population* Washington Government Printing Office, various

U S. Department of Transportation *Transportation Planning Data for Urban Areas.* Washington Government Printing Office, 1971, 1973

The Incidence and Efficiency Effects of Taxes on Income from Capital

---

John B. Shoven

Stanford University

This article reexamines the incidence and efficiency cost of the discriminatory taxation of capital income in the United States It is argued that Harberger's 1966 estimates of the static welfare loss were subject to two important mistakes Their correction lowers the efficiency cost estimates approximately 38 percent The paper also compares the corrected results of the Harberger model with those achieved with an algorithmic solution procedure for a general equilibrium model When the latter approach is used with the same two-sector division of production, the results are very similar to those of Harberger's model. With disaggregation to 12 production sectors, however, the loss estimates increase by an average of 40 percent

I. Introduction

There has developed over the past 15 years or so a very extensive literature regarding the taxation of income from capital, particularly the corporation income tax On issues regarding the distributional incidence and inefficiency of these capital income levies, the works of Arnold Harberger (1959, 1962, 1966) have been particularly influential and important. His models and results often serve as a reference with which other approaches to these questions are compared The 1959 and 1966 articles by Harberger deal with estimating the efficiency cost or dead-weight loss of these taxes resulting from the misallocation of resources The latter of these pieces is somewhat more complete, in that it attempts to include all taxes on capital income rather than just the corporation income tax. Harberger's 1962 article describes his theoretical framework and addresses the question

This work was supported by National Science Foundation grant GI-39319 at the Institute for Mathematical Studies in the Social Sciences, Stanford University The author wishes to thank J. G Ballentine, John Whalley, and Mark Gertler for helpful comments.

[*Journal of Political Economy*, 1976, vol. 84, no. 6]

© 1976 by The University of Chicago. All rights reserved.

of the incidence of the corporation income tax on the functional distribution of income. On this issue, Harberger concludes that capital bears close to the full burden of the tax for realistic estimates of the elasticity parameters of his model.

Despite the attention economists have given the corporation income tax and other capital income taxes, nothing near a consensus has emerged regarding their economic impact. The importance of the subject is often rejuvenated by suggestions that the corporation income tax be integrated into the personal income tax structure.

This article has two primary purposes with respect to capital income taxation. The first of these is to correct two serious flaws in Harberger's 1966 article. One error is purely arithmetic, while the other is conceptual, but they both significantly affect his widely referred to inefficiency estimates. In the second section of this paper, I briefly describe the Harberger model,[1] comment on its shortcomings, illustrate the sources of the two mistakes, and present the corrected results for the Harberger approach. The importance of this section derives from the fact that the corrected inefficiency estimates of Harberger's model amount to only 32–63 percent of the published figures for the same elasticity parameterizations. On the other hand, it will be argued that the usual reporting of these dead-weight loss results as a trivial fraction of GNP is misleading. The section is also of value in displaying how important consistent units definitions are in applying data to the Harberger model.

The second purpose of this paper is to compare the corrected results of the Harberger model with those derived from an algorithmic solution of a general equilibrium model. One advantage of such an approach is that, in not using calculus, the changes which one is trying to gauge need not be assumed small in the model itself. The task of comparing the results of such a procedure with those of the Harberger model was undertaken in an earlier work (Shoven and Whalley 1972), but a recalculation is now in order, for several reasons. First, the earlier article used the data and the results of the Harberger 1966 article in evaluating the two techniques. However, the data and therefore the results were incorrect (as shown here in Sec. II). Second, the algorithmic approach has been greatly improved in the interim, as is detailed in Shoven and Whalley (1973) and Shoven (1976). And finally, third, another advantage of the algorithmic approach is that it is capable of evaluating an *n* sector rather than a two-sector model. Therefore, in order to provide a feel for the effects of aggregated analysis, the impact of the uneven capital income tax levies is estimated for a 12- as well as a two-sector breakdown of the production side of the U.S. economy. With the algorithmic approach, the flat tax rate necessary to generate the same real yield as the current system is calculated.

---

[1] The model has been elaborately specified several times in the literature. See, e.g., Harberger (1962); appendix B of Shoven and Whalley (1972); McLure (1974), or McLure (1975).

II. Harberger Model, Data, and Results

The Harberger model, as detailed in his 1962 article and in several subsequent studies, is sufficiently well known and documented that it need not be completely specified here However, since a primary purpose of this section is to correct the results of his 1966 paper concerning the efficiency losses of the taxes, it is appropriate that we review the approach of that study.

In his model, Harberger makes an empirical distinction between a heavily and lightly taxed sector These sectors are sometimes referred to as the "corporate" and "non-corporate" sectors, due to the major role played by the corporation income tax in causing the differential rates, although the sectoral division does not exactly correspond to the legal distinction between incorporated and unincorporated enterprises. Harberger assumes that each sector employs two factors, capital services and labor, in the production of homogeneous outputs

In order to estimate the efficiency loss due to the differential taxation of the return to capital, Harberger applies a form of welfare analysis in the tradition of Marshallian producer surplus It is assumed that the marginal product of capital schedules for each sector are linear, as drawn in figure 1. Output units are chosen so that both commodity prices are unity, and therefore the schedules in figure 1 can be thought of as marginal revenue product schedules The total quantity of both capital and labor in the economy is assumed fixed and always fully employed. With these assumptions, the changes in the capital allocation can be used to generate a measure of the social waste imposed by the distortion. In the absence of any taxes, capital will allocate itself in a market economy such that the rate of return $\bar{r}$ is equal for the two sectors and the capital endowment is fully employed Upon the imposition of a tax on capital income in sector X, the gross rate of return r_g in that sector must be such that the net rate of return r_n is equalized across the sectors and capital is again fully employed. The difference between r_g and r_n is by definition the tax T per unit of capital utilized in sector X.

Referring to figure 1, the area $ABEF$ can be interpreted as the value of the lost output in sector X when K_X decreases from K_{X0} to K_{X1} upon the imposition of the tax Analogously, $GHIJ$ is the value of increase in output in sector Y. Since we know that capital is fully employed both in the presence and absence of the tax, it must be true that $K_{X0} - K_{X1} = K_{Y1} - K_{Y0}$. The area $FECD$ represents the social loss of the tax in the producer surplus sense (it is simply $ABEF - GHIJ$) and is given by

$$\tfrac{1}{2}(r_g - \bar{r})(K_{X0} - K_{X1}) + \tfrac{1}{2}(\bar{r} - r_n)(K_{Y1} - K_{Y0}) = -\tfrac{1}{2}T\Delta K_X, \quad (1)$$

where $\Delta K_X = K_{X1} - K_{X0} = K_{Y0} - K_{Y1}$ and $T = r_g - r_n$.

Now, in order to estimate the magnitude of the social efficiency loss, an expression for ΔK_X is required In order to obtain such a solution,

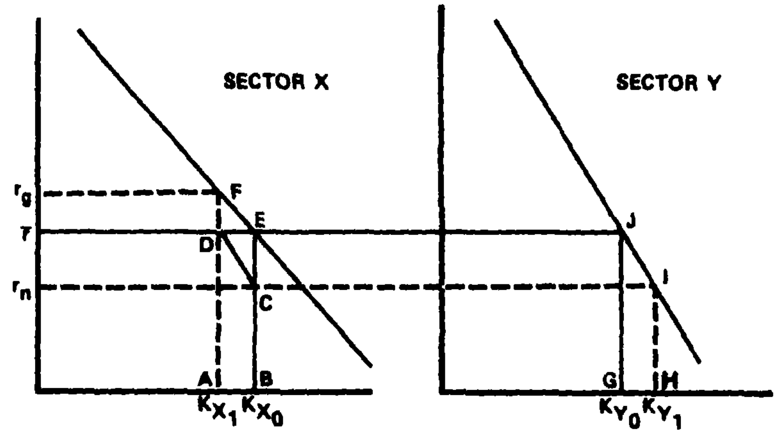


Fig 1

Harberger calls upon the static two-sector, two-factor, general equilibrium model of his 1962 paper. The use of this model in conjunction with the Marshallian approach above has several shortcomings, particularly if the purpose involves measurement of economic changes. The primary problem is that local or "small change" assumptions are made repeatedly in the analysis For example, in solving for the efficiency loss formula (eq [1]), it has been assumed that the marginal product of capital varies linearly with the amount of capital for each of the sectors. In addition, either the labor allocation must be assumed to be unaltered, or the marginal product of capital must be taken as independent of the amount of labor employed in each sector. Without either of these conditions, the marginal product curves themselves will not be stationary. At the same time, in the model used to solve for the capital shift, ΔK_X, constant returns to scale in production is assumed. The joint assumptions of linear marginal product schedules, constant returns to scale, and separable marginal products (or an unchanged labor allocation) are inconsistent for large changes. Given that the purpose of the analysis is to determine the magnitude of the effects of distortionary taxation, assumptions such as these which are consistent only locally are undesirable.

Solving through the equations of the Harberger (1962) model, the following solution for ΔK_X is obtained:

$$\Delta K_X = \frac{K_X T\{-E[g_K S_X(L_X/L_Y) + f_K S_Y] - S_X S_Y f_L\}}{E(g_K - f_K)[(K_X/K_Y) - (L_X/L_Y)] - S_Y - S_X[f_L(K_X/K_Y) + f_K(L_X/L_Y)]}, \tag{2}$$

where E = compensated price elasticity of demand for X; $S_X(S_Y)$ = elasticity of factor substitution in sector $X(Y)$; $f_K(g_K)$ = share of capital in sector $X(Y)$; and $f_L(g_L)$ = share of labor in sector $X(Y)$ So, as one would expect, the capital shift depends on the various elasticities and factor intensities and is also proportional to T. This last point is emphasized, as it later becomes important in that it, along with equation (1), implies that the social cost varies with the square of the tax rate

Harberger similarly solves this system of equations for the change in the net price of capital (i.e , $\Delta P_K = r_n - \bar{r}$) and obtains

$$\Delta P_K = \frac{T\{Ef_K[(K_X/K_Y) - (L_X/L_Y)] + S_X[f_L(K_X/K_Y) + f_K(L_X/L_Y)]\}}{E(g_K - f_K)[(K_X/K_Y) - (L_X/L_Y)] - S_Y - S_X[f_L(K_X/K_Y) + f_K(L_X/L_Y)]}. \tag{3}$$

He uses this figure to evaluate the incidence of the burden, reasoning that, if $\Delta P_K = -(TK_X)/(K_X + K_Y)$, capital could be said to bear the full burden of the tax in that its gross return would be unchanged, while if $\Delta P_K = 0$ both net input prices would be unaffected, and thus the share of national income going to capital and labor would remain constant.

In order to evaluate expressions (2) and (3) for the 1953–59 U.S economy, Harberger uses and supplements the data of Rosenberg (1969) However, in carrying this out, two mistakes are made. Estimates of K_X, K_Y, and T are drawn from table 1, reproduced from Harberger (1966) Columns 1 and 2 are derived from Rosenberg's disaggregated study, while column 3 is meant to reflect the personal income taxes on capital income. Appealing to columns 4 and 5, Harberger notes that total taxes on net income in the "non-corporate" and "corporate" sectors average, respectively, 45 percent and 168 percent [2] Thus he asserts that the taxation of income from capital in the United States during this period may be approximated by a general tax of 45 percent on all net income from capital and an 85 percent surtax on the net income from capital originating in the heavily taxed sector (1 45 × 1.85 = 2 68).

The first flaw in Harberger's 1966 application of this data to his model is an inconsistency between the calculation of tax rates above and the units of measurement of capital chosen in the study Harberger takes as one unit of capital that amount which earned an average annual flow of $1.00 net of all taxes. Thus, appealing to column 5 of table 1, he sets K_Y as 18,510 million and K_X as 19,547 million In models such as this, with a fixed quantity of aggregate capital, a general tax on all capital income is nondistortionary Therefore Harberger proceeds to evaluate only the 85 percent surtax applying to the capital income of the "cor-

---

[2] The use of average tax data to evaluate distortions has been questioned by Stiglitz (1975). The implied assumption that the marginal gross cost of capital equals the average gross cost can be defended on the grounds that investment decisions are often made using average capital costs (see, e.g , Van Horne 1971)

TABLE 1

TAXES ON INCOME FROM CAPITAL, BY MAJOR SECTORS
(Annual Averages, 1953–59, in $ Millions)
(Reproduced from Harberger 1966, p. 110)

	Total Income[a] from Capital (1)	Property & Corp. Income Taxes (2)	Other Tax Adjustments (3)	Total Tax on Income from Capital (4)	Net Income from Capital (5)
"Non-corporate" sector	26,873	6,639	1,724	8,363	18,510
Agriculture	7,481	1,302	927[a]	2,229	5,252
Housing	18,429	5,140	797[b]	5,937	12,492
Crude Oil and Gas	963	197	[c]	197	766
"Corporate" sector	52,399	22,907	9,945[d]	32,852	19,547
Total	79,272	29,546	11,669	41,215	38,057

[a] "Income" (Rosenberg 1969, p. 125) is defined as income from capital for nonfinancial industry and includes

(1) Corporate sector net income before corporate profits tax liability and property tax payments
(2) For the unincorporated sector, the portion of the total income of the unincorporated enterprise that is a return on equity capital, plus property tax payments
(3) Net monetary interest paid by businesses on borrowed capital in the form of debt obligations
(4) Net rent paid by an industry to persons for the use of physical capital
(5) Net realized capital gains by the corporate sector that are considered as income to an industry

[a] Assumes a 15% effective tax on income from capital in agriculture after payment of property and corporate income taxes (i.e. $[3]^a = 15\%$ of $[1] - [2]$)

[b] Assumes that 70% of income from capital in the housing sector is generated by owner-occupied housing, on which no personal income tax liability is incurred. It is assumed that the remaining 30% of capital income from housing is subjected to a 20% income tax rate after the deduction of property and corporate income taxes incurred (i.e., $[3]^b = 6\%$ of $[1] - [2]$)

[c] Assumes personal tax offsets on account of oil depletion allowances and similar privileges offset any taxes on dividends and capital gains in this sector

[d] Assumes a 50% dividend distribution rate and a "typical" effective tax rate of 40% on dividend income (i.e., $[3]^d = 20\%$ of $[1] - [2]$)

porate" sector (i.e., he set $T = 0.85$ in equations (1), (2), and (3)). The problem is that the surtax is in fact applied not to the net capital income of the heavily taxed sector (indicated in table 1 to be 19,547 *million) but to the capital income in that sector gross of the neutral* 45 percent tax ($28,379 million). Harberger is thus evaluating a smaller tax than his data reveals! The social marginal product of each of Harberger's units of capital is $1.45, not the $1.00 he recognizes. In order to rectify this problem, one either has to evaluate ***both*** the 45 percent and 168 percent taxes simultaneously or to evaluate only the surtax and take as a unit of capital that amount which earns $1.00 net of the surtax but gross of the neutral tax. The dead-weight loss estimates are identical with either convention, and I have chosen the latter simply because the previous figure and equations have been for the surtax case. Under thi approach, table 1 shows $K_Y = 26{,}873$ million and $K_X = 28{,}379$ million. As the estimate of the capital shift and the dead-weight loss are pro portional to K_X, this correction alone would increase Harberger's in efficiency results 45 percent.

Unfortunately, the problem with units above is not the most serio error related to the data of Harberger (1966). They contain a simp arithmetic mistake, which also greatly affects his results. The entry superscripted d in column 3 should be 5,898 rather than 9,945, and, with this correction, the net income from capital in the "corporate" sector becomes \$23,589 rather than \$19,547 million The total tax on capital income in the "corporate" sector should be \$28,805 million The source of the error is obvious In determining the column 3 entry for the "corporate" sector, the arithmetic was mistakenly done on the first two entries in the total row. Note that 9,945 is 20 percent of 79,272 less 29,546.[3]

Table 2 contains the corrected factor input and capital tax data for both a two- and a 12-sector level of aggregation for averaged data for the same 1953–59 period The capital income and tax data are again drawn from Rosenberg (1969), and the "Other Capital Tax Adjustments" column was constructed upon the assumptions listed in the footnotes to table 1 The labor data are derived from National Income and Product Accounts At the two-sector level of aggregation, the main difference between the two tables is caused, of course, by the arithmetic mistake just noted The surtax rate, which Harberger figures as 85 percent, is, according to table 2, only 53 percent ($1\ 45 \times 1\ 53 = 2\ 22$). The effect of this error is magnified, since, as was pointed out earlier, the loss estimate varies as the square of the rate of distortionary taxation. The capital data, using the consistent definition of units mentioned above, now indicates that $K_Y = 26{,}878$ million and $K_X = 34{,}244$ million.

Subject to these two mistakes, Harberger uses the data of table 1 augmented with labor data[4] to estimate ΔK_X and $-(1/2)T\Delta K_X$ for several different elasticities of factor substitution (S_X and S_Y) and for two different compensated elasticities of demand for the "corporate" product X. The two demand elasticities E are based on the elasticity of substitution between X and Y, which is labeled V, being assigned values of 1 and $\frac{1}{2}$ The relationship between V and the compensated demand for X is

$$E = Vr_Y, \tag{4}$$

where r_Y is the share of national income spent on Y. Thus the $V = 1$ case is consistent with demands derived from a Cobb-Douglas-type utility function. Harberger takes r_Y as 0 17, although my revised data (col. 3, table 2) suggests that r_Y is just slightly under 0.15

Many different estimates have been made of the elasticities of substitution between labor and capital for various sectors, often concluding with contradictory results. Most of the work based on cross-section data

---

[3] There are two other errors in table 1 The total income figure from capital in the noncorporate sector should be 26,878 rather than 26,873, and in the corporate sector the figure should be 52,394 rather than 52,399 The total of the two is correct in table 1.

[4] Harberger takes L_X to be \$200 billion and L_Y to be \$20 billion Col 1 of table 2 agrees very closely with his L_X figure but shows L_Y to be only \$17,471 billion (or units).

TABLE 2

FACTOR PAYMENTS AND CAPITAL INCOME TAXES, BY MAJOR SECTORS
(Annual Averages, 1953–59, in $ Million)

Sector	Total Return to Labor (1)	Total Return to Capital (2)	Total Factor Return (1) + (2) (3)	Property and Corporate Income Tax (4)	Other Capital Tax Adjustments (5)	Total Capital Tax (4) + (5) (6)	Net Return to Capital (2) − (6) (7)	Capital Tax as a Fraction of Total Return to Capital (6)/(2) (8)	Capital Tax as a Fraction of Net Return to Capital (6)/(7) (9)
"Non-corporate"	17,471	26,878	44,349	6,639	1,724	8,363	18,515	0.3111466	0.4516878
Agriculture	8,800	7,481	16,281	1,302	927	2,229	5,252	0.2979548	0.4244097
Real estate	6,869	18,429	25,298	5,140	797	5,937	12,492	0.3221552	0.4752641
Crude oil and gas	1,802	968	2,770	197		197	771	0.2035123	0.2555123
"Corporate"	199,871	52,394	252,265	22,907	5,898	28,805	23,589	0.5497766	1.2211220
Mining*	2,528	688	3,216	305	77	382	306	0.5552325	1.2483660
Contract construction	16,670	1,195	17,865	435	152	587	608	0.4912133	0.9654605
Manufacturing†	79,626	24,665	104,291	12,488	2,435	14,923	9,742	0.6050273	1.5318209
Lumber and wood products	2,426	718	3,144	206	102	308	410	0.4289693	0.7512195
Petroleum and coal products	1,846	3,028	4,874	770	452	1,222	1,806	0.4035667	0.6766334
Trade	43,590	10,897	54,487	3,493	1,481	4,974	5,923	0.4564559	0.8397771
Transportation	14,078	2,683	16,761	1,230	291	1,521	1,162	0.5669027	1.3089500
Communication and public utilities	7,394	6,489	13,883	3,290	640	3,930	2,559	0.6056403	1.5357561
Services	31,713	2,031	33,744	690	268	958	1,073	0.4716888	0.8928238
Total	217,342	79,272	296,614	29,546	7,622	37,168	42,104	0.4688666	0.8827664

* Other than crude oil and gas.

† Other than lumber and wood products and petroleum and coal products.

suggests that, for most two-digit manufacturing industries, S is not significantly different from 1 (see, e.g., Solow 1964 and Minasian 1961). On the other hand, time-series studies yield estimates significantly less than 1 (see Lucas 1969). Given this situation, Harberger looks at several combinations of S_X and S_Y.

Table 3 contains Harberger's published results, appropriately qualified as only rough estimates, and the impact on them of the two errors just discussed. Their importance is readily apparent, although the net effect is somewhat reduced, as they cut in opposing directions. The 45 percent increase in the inefficiency estimates due to the units-definition problem is apparent in comparing column 7 with column 5. The independent effect of the arithmetic mistake and Harberger's use of other data which are only roughly correct is seen by contrasting columns 9 and 7. The range of the estimates is lowered from the published \$1.0 billion–\$2.9 billion to the \$0.625 billion–\$1.79 billion shown in column 9 of table 3 and based on the data of table 2. Surprisingly, Harberger's numerical estimates of the outflow of capital from the "corporate" sector due to the discriminatory taxation shown in column 4 are almost identical with the corrected figures given in column 8 except in case II, where he has made a serious computational error. The equivalency is less comforting and more coincidental, however, when it is recalled that each of Harberger's units of capital is equivalent to 1.45 units under the revised system of accounts.

It may be claimed that all that has been done is to lower further Harberger's already small loss estimates, widely reported as between 0.5 and 1 percent of GNP (or NNP). This presentation of the results as a percentage of national product seems misleading, however. Perhaps a more natural benchmark is the net revenue yield of the surtax. This alternative, also not without some interpretational difficulties, suggests a "coefficient of inefficiency" which can be defined as the ratio of the deadweight loss of the surtax (in this case the 53% additional tax paid by sector X) to the net surtax revenue. This additional revenue due to the presence of the surtax amounts to between \$9.3 billion and \$13.9 billion, depending on the factor-elasticity assumptions. The yield of a flat 45 percent tax on net income depends upon the net price of capital in the nondistortionary situation. For the revised results of column 9, table 3, the coefficient of inefficiency ranges from 0.06 to 0.15. Column 10 of table 3 records the values for ΔP_K, the change in the price of capital in moving from the nondistortionary to the distortionary situation. This variable sheds some light on the functional incidence of the taxation, particularly when it is noted that $-TK_X/(K_X + K_Y) = -0.297$. This is the value of ΔP_K at which capital can be said to bear the entire tax, as its net return is reduced by the amount of the tax proceeds. For those cases in which the price falls by more than this figure, capital can be said to be bearing more than 100 percent of the tax burden.

TABLE 2

FACTOR PAYMENTS AND CAPITAL INCOME TAXES, BY MAJOR SECTORS
(Annual Averages, 1953–59, in $ Million)

Sector	Total Return to Labor (1)	Total Return to Capital (2)	Total Factor Return (1) + (2) (3)	Property and Corporate Income Tax (4)	Other Capital Tax Adjustments (5)	Total Capital Tax (4) + (5) (6)	Net Return to Capital (2) − (6) (7)	Capital Tax as a Fraction of Total Return to Capital (6)/(2) (8)	Capital Tax as a Fraction of Net Return to Capital (6)/(7) (9)
"Non-corporate"	17,471	26,878	44,349	6,639	1,724	8,363	18,515	0.3111466	0.4516878
Agriculture	8,800	7,481	16,281	1,302	927	2,229	5,252	0.2979548	0.4244097
Real estate	6,869	18,429	25,298	5,140	797	5,937	12,492	0.3221552	0.4752641
Crude oil and gas	1,802	968	2,770	197		197	771	0.2035123	0.2555123
"Corporate"	199,871	52,394	252,265	22,907	5,898	28,805	23,589	0.5497766	1.2211220
Mining*	2,528	688	3,216	305	77	382	306	0.5552325	1.2483660
Contract construction	16,670	1,195	17,865	435	152	587	608	0.4912133	0.9654605
Manufacturing†	79,626	24,665	104,291	12,488	2,435	14,923	9,742	0.6050273	1.5318209
Lumber and wood products	2,426	718	3,144	206	102	308	410	0.4289693	0.7512195
Petroleum and coal products	1,846	3,028	4,874	770	452	1,222	1,806	0.4035667	0.6766334
Trade	43,590	10,897	54,487	3,493	1,481	4,974	5,923	0.4564559	0.8397771
Transportation	14,078	2,683	16,761	1,230	291	1,521	1,162	0.5669027	1.3089500
Communication and public utilities	7,394	6,489	13,883	3,290	640	3,930	2,559	0.6056403	1.5357561
Services	31,713	2,031	33,744	690	268	958	1,073	0.4716888	0.8928238
Total	217,342	79,272	296,614	29,546	7,622	37,168	42,104	0.4688666	0.8827664

---

* Other than crude oil and gas.

† Other than lumber and wood products and petroleum and coal products.

suggests that, for most two-digit manufacturing industries, S is not significantly different from 1 (see, e g , Solow 1964 and Minasian 1961) On the other hand, time-series studies yield estimates significantly less than 1 (see Lucas 1969) Given this situation, Harberger looks at several combinations of S_X and S_Y

Table 3 contains Harberger's published results, appropriately qualified as only rough estimates, and the impact on them of the two errors just discussed Their importance is readily apparent, although the net effect is somewhat reduced, as they cut in opposing directions The 45 percent increase in the inefficiency estimates due to the units-definition problem is apparent in comparing column 7 with column 5 The independent effect of the arithmetic mistake and Harberger's use of other data which are only roughly correct is seen by contrasting columns 9 and 7 The range of the estimates is lowered from the published \$1 0 billion–\$2 9 billion to the \$0 625 billion–\$1 79 billion shown in column 9 of table 3 and based on the data of table 2 Surprisingly, Harberger's numerical estimates of the outflow of capital from the "corporate" sector due to the discriminatory taxation shown in column 4 are almost identical with the corrected figures given in column 8 except in case 11, where he has made a serious computational error The equivalency is less comforting and more coincidental, however, when it is recalled that each of Harberger's units of capital is equivalent to 1 45 units under the revised system of accounts.

It may be claimed that all that has been done is to lower further Harberger's already small loss estimates, widely reported as between 0 5 and 1 percent of GNP (or NNP) This presentation of the results as a percentage of national product seems misleading, however Perhaps a more natural benchmark is the net revenue yield of the surtax This alternative, also not without some interpretational difficulties, suggests a "coefficient of inefficiency" which can be defined as the ratio of the deadweight loss of the surtax (in this case the 53% additional tax paid by sector X) to the net surtax revenue. This additional revenue due to the presence of the surtax amounts to between \$9 3 billion and \$13 9 billion, depending on the factor-elasticity assumptions The yield of a flat 45 percent tax on net income depends upon the net price of capital in the nondistortionary situation For the revised results of column 9, table 3, the coefficient of inefficiency ranges from 0 06 to 0 15 Column 10 of table 3 records the values for ΔP_K, the change in the price of capital in moving from the nondistortionary to the distortionary situation This variable sheds some light on the functional incidence of the taxation, particularly when it is noted that $-TK_X/(K_X + K_Y) = -0\ 297$. This is the value of ΔP_K at which capital can be said to bear the entire tax, as its net return is reduced by the amount of the tax proceeds For those cases in which the price falls by more than this figure, capital can be said to be bearing more than 100 percent of the tax burden.

TABLE 3

ESTIMATES OF EFFICIENCY COST OF EXISTING TAXES ON INCOME FROM CAPITAL (1953–59), USING HARBERGER'S MODEL

CASE	S_X (1)	S_Y (2)	V (3)	HARBERGER'S ESTIMATES* ΔK_X (4)§	HARBERGER'S ESTIMATES* $-(1/2)T\Delta K_X$ (5)‖	ESTIMATES WITH REVISED UNITS† $\Delta K_X'$ (6)§	ESTIMATES WITH REVISED UNITS† $-(1/2)T\Delta K_X'$ (7)‖	ESTIMATES WITH REVISED UNITS AND REVISED DATA‡ $\Delta K_X''$ (8)§	ESTIMATES WITH REVISED UNITS AND REVISED DATA‡ $-(1/2)T\Delta K_X''$ (9)‖	ESTIMATES WITH REVISED UNITS AND REVISED DATA‡ ΔP_K (10)
1	−1.0	−1.0	−1.0	−6.9	2.9	−9.7	4.1	−6.76	1.79	−0.326
2	−1.0	−0.5	−1.0	−5.9	2.5	−8.4	3.6	−5.85	1.55	−0.362
3	−0.5	−1.0	−1.0	−5.2	2.2	−7.4	3.1	−5.20	1.38	−0.221
4	−0.5	−0.5	−1.0	−4.8	2.0	−6.7	2.9	−4.75	1.26	−0.261
5	−1.0	0.0	−1.0	−4.7	2.0	−6.6	2.8	−4.71	1.25	−0.408
6	−0.5	0.0	−1.0	−3.9	1.7	−5.7	2.4	−4.09	1.08	−0.318
7	−1.0	−1.0	−0.5	−5.3	2.3	−7.5	3.2	−5.13	1.36	−0.365
8	−1.0	−0.5	−0.5	−4.2	1.8	−5.8	2.5	−4.00	1.06	−0.408
9	−0.5	−1.0	−0.5	−4.1	1.7	−5.8	2.5	−4.06	1.08	−0.272
10	−0.5	−0.5	−0.5	−3.5	1.5	−4.9	2.1	−3.38	0.896	−0.326
11	−1.0	0.0	−0.5	−5.0	2.1	−3.6	1.5	−2.55	0.676	−0.464
12	−0.5	0.0	−0.5	−2.4	1.0	−3.3	1.4	−2.36	0.625	−0.408

NOTE.—$S_X(S_Y)$ = elasticity of factor substitution in "corporate" ("non-corporate") sector, V = elasticity of substitution between products X and Y

* Based on $K_X = 20$, $K_Y = 20$, $L_X = 200$, $L_Y = 20$, $f_K = 0.2$, $f_L = 0.8$, $g_K = 0.54$, $r_Y = 0.17$, $T = 0.85$ Harberger made a computational error in case 11, where he should have gotten $\Delta K_X = -2.5$ and $-(1/2)T\Delta K_X = 1.1$

† Based on $K_X = 28.5$, $K_Y = 27.0$, $L_X = 200$, $L_Y = 20$, $f_K = 0.2$, $f_L = 0.8$, $g_K = 0.54$, $r_Y = 0.17$, $T = 0.85$

‡ Based on $K_X = 34.244$, $K_Y = 26.878$, $L_X = 199.871$, $L_Y = 17.471$, $f_K = 0.20769$, $f_L = 0.79231$, $g_K = 0.60606$, $r_y = 0.14952$, $T = 0.53003$

§ In billions.

‖ In $ billion.

III. Results of the Algorithmic Approach

Using the revised data of table 2, the efficiency cost of the distortionary taxation of capital income has been recomputed using a solution algorithm for a general equilibrium model It is, of course, the corrected results of columns 9 and 10 of table 3 which will serve as a basis for comparison. Such additional questions as the impact of the taxes on the supply of labor, on the functional and personal distribution of income, and the flat capital income tax rate which would be required to generate the same real revenue as the existing taxes can be addressed using the algorithmic approach In most cases the assumptions of Harberger have been followed, although the models of course are not strictly comparable Production functions are assumed to be of the CES form, while both CES and Cobb-Douglas utility functions were analyzed for consumers Two classes of consumers were incorporated, one representing the top 10 percent of the income recipients in the United States, the other the bottom 90 percent The model was solved both with and without the inclusion of a labor-leisure choice, although detailed results will be presented only for the fixed-labor-supply case

Just as there is no need to respecify the Harberger model completely each time it is employed, it is likewise unnecessary to review fully the Scarf (1973) general equilibrium algorithm utilized here to derive the results of this section For that reason, only a brief sketch of the mechanics of the approach is presented

The basic idea of the algorithmic approach of this section is to specify an Arrow-Debreu general equilibrium model with and without taxes The existence of an equilibrium for a model with such ad valorem taxes is assured in Shoven (1974) In order to parameterize the model, some extraneous estimates are required These are the same factor substitution and demand elasticities which must be specified in the Harberger approach described in the last section Given these estimates, which comprise perhaps one-third of the total number of parameters, the model is solved "backward" to determine the values that the remaining variables must take in order to replicate exactly the prechange economy The technique of carrying out this parameterization approach is more fully described in Shoven (1973) and Whalley (1973) Once all the parameters are set, the model is solved "forward" for a new equilibrium in the presence of the new tax environment In the case of this study, the new system involves the removal of the surtax on capital income in the heavily taxed sector (or sectors with the 12-sector level of aggregation). The solution technique is a version of Scarf's algorithm modified to include taxes and continuous production functions. This method is detailed in Shoven (1973) and Shoven and Whalley (1973). The degree of approximation is greatly improved with a Newton-type termination routine similar to that described in appendix A of Shoven and Whalley (1972). The advantages of this general approach over that of Section II are that (1) no localization assumptions are required; (2) it is simple to incorporate many commodities and several classes of consumers; and (3) it is unnecessary to assume fixed factor supplies. A more abstract advantage is that the model need not be kept sufficiently simple so as to be analytically tractable

On the production side of the economy, each sector's technological production possibilities are characterized by a CES production function such as

$$Q_i = \gamma_i[\alpha_i L_i^{-\rho_i} + (1 - \alpha_i)K_i^{-\rho_i}]^{-1/\rho_i} \tag{5}$$

Noting the discussion in the previous section of plausible estimates of the elasticity of substitution (in this case $S_i = -1/[1 + \rho_i]$), seven cases were considered: three in line with Harberger, and four others They may be listed as

Case	S_X	S_Y
1	−1.00	−1.00
2	−1.00	−0.50
3	−1.00	−0.25
4	−0.75	−0.25
5	−0.50	−0.50
6	−0.50	−0.25
7	−0.25	−0.25

As previously mentioned, two consumers were considered; the first heuristically represents the upper 10 percent of the income recipients, while the second represents the lower 90 percent The first is endowed with approximately 23 percent of the economy's labor (corresponding to the observed share of labor income going to the top decile of income receivers) and 40 percent of total stock of capital (both figures from Projector and Weiss 1966) This latter figure roughly corresponds to the share of capital income going to the top 10 percent of the income receivers, although it is much lower than the share of capital income going to the top 10 percent of wealth holders. Endowing the high-income receivers with more than 10 percent of the labor appeals to an equal endowment of labor in natural units but a disproportionate endowment in efficiency units.

Each consumer's demand functions are of the form

$$x_{ij} = \frac{I_j}{\sum_{i=1}^{n} a_{ij} P_i^{(1-V_j)}} \frac{a_{ij}}{P_i^{V_j}}, \qquad i = 1, \ldots, n; j = 1, 2, \tag{6}$$

where x_{ij} is consumer j's demand for commodity i, a_{ij} measures the intensity of his desire for commodity i, P_i is the price of the ith good, V_j

is individual j's elasticity of substitution between commodities, and I_j is his income given by

$$I_j = \sum_{i=1}^{n} P_i w_{ij}, \tag{7}$$

where the w_{ij} are his initial asset holdings, including labor.

In the recomputations, two values for the consumer's elasticities of substitution (i.e., the V_j's) are examined: 1.0 and 0.5. With V_j equal to 1.0, the demand functions (6) are of the familiar Cobb-Douglas type.

Using the demand functions (6) permits one to impose the observed aggregate 5.69–1.0 expenditure ratio for the outputs of the heavily and lightly taxed sectors. Approximately \$252 billion a year was spent on "corporate" products, while only \$44 billion was spent on the output of the "non-corporate" sector. If each individual's tastes were such that $a_{1j} = 5.69a_{2j}$, $j = 1, 2$, where the "corporate" product is labeled commodity 1 and the "non-corporate" product commodity 2, the expenditure ratios of the model would exactly correspond to the 5.69–1.00 figure with unitary prices. However, in the examples investigated here, lower-income individuals are taken to place a relatively higher weight on the output of the lightly taxed sector than higher-income individuals. This is consistent with observed higher budget shares allocated to food expenditures (agricultural output) by lower-income people. The ratios a_{1j}/a_{2j} used are 7.00 for the higher-income consumer and 5.30 for the lower-income consumer.

The total endowments of the economy are taken from table 2 of the last section. Again, we need to clarify units conventions. The approach taken in this section is to analyze all capital income taxes, and not simply the surtax. This implies taking as a unit of capital that amount which earns \$1.00 net of all taxes, Harberger's original units. The advantage of this convention here is that it readily permits disaggregation, and it explicitly recognizes that the government's tax collections rise (even from neutral taxes) when the net price of capital services increase. It is a perfectly acceptable units convention, provided all taxes are incorporated into the model. Relative price changes and real output effects are independent of which of the two consistent units definitions is chosen. Only in comparing absolute quantities of capital must the appropriate conversion (1.45) be applied. With this settled and in the absence of a labor-leisure choice, the total labor endowment is taken to be 217.342 billion units, while the capital-services endowment is 42.104 billion units. In the presence of the surtax, 23.589 billion units of capital and 199.871 billion units of labor are allocated to the "corporate" sector (from table 2, columns 1 and 7). To focus solely on the impact of the tax, the fraction of government tax proceeds returned to each consumption class is assumed equal to that class's fraction of the total capital endowment.

Thus the potential redistributionary effects of the expenditure of the tax proceeds are not considered here.

Table 4 contains a summary of the efficiency loss and incidence results for the two production-sector cases What is compared in this table is the observed equilibrium (the equilibrium in the presence of distortionary capital income taxes) with the equilibrium which would prevail in the absence of the distortionary taxes (i.e , in the absence of what Harberger refers to as the surtax) Also reported are the equilibrium prices and the tax rate (on net income) which would prevail if the replacement were a flat tax designed to have the same real yield as is realized in the distortionary situation. A Paasche price index is used in determining real rather than nominal revenue equivalency. The technique of determining this equal-yield tax rate is described in Shoven and Whalley (1975) The entry termed the "shift factor" is presented to give some insight into the incidence of the distortionary capital taxation with respect to the functional distribution of income [5] The definition of the term is shift factor = $1 + (\Delta P_K K)/(\Delta R)$, where K is the total endowment of capital, ΔP_K is the change in the net price of capital in moving from the nondistortionary to the distortionary tax regime (thus this term is negative in all cases), and ΔR is the change in nominal government revenue (which is of course positive) If this shift factor is zero, the decrease in the net return to capital is equal to the increase in government revenue due to the distortionary surtax, and in this sense capital may be said to bear the full burden A negative value for this shift factor would indicate that capital bears more than 100 percent of the burden of the surtax, while a positive value would reflect the fact that labor is sharing in its costs.

Table 4, part A, presents the results for the cases in which both consumer classes have demand elasticities of substitution (V_j's) equal to 1 As one would expect, the removal of the surtax involves a decrease in the price of X, the "corporate" output, and an increase in the price of Y (due to the increase in the net price of capital) It might be noted that these price changes tend to benefit the higher-income consumer class more than the lower-income class, as the former spends a larger fraction of its income on products which decrease in price (i.e , X) As is done in Harberger's work, labor is taken as the numéraire commodity and thus always has a price of unity. The "X" and "Y" rows show that the output of sector X increases by 2–3 percent, while the output of sector Y decreases by as much as 15 percent. Factor substitution occurs, and the "corporate" sector switches to a more capital-intensive technology with the removal of the surtax, while the lightly taxed sector becomes more labor intensive. One interesting aspect of this particular problem is that the sector whose capital income is currently heavily taxed is relatively labor intensive. This fact opens the possibility that labor may bear a large fraction of the burden of the surtax, particularly in cases with low elasticities of factor substitution. Indeed, the "shift factor" row of this section shows that, in cases 5, 6, and 7, labor's share of the burden is significant. In fact, in case 7 the relative share of the rich consumer class (whose assets are capital intensive) is 0.4 percent higher in the presence of the distortionary tax than in its absence. The efficiency loss estimates are shown to be sensitive to the specification of the production parameters and range between $0.435 billion and $2.344 billion. These figures amount to 2.8–18.7 percent of the revenue generated by the surtax. For each parameterization, two loss estimates (or changes in NNP)[6] are given: one evaluated at the observed prices and one at the new, nondistortionary prices. As is known from index number theory, under "reasonable" assumptions it can be shown that the estimate based on the new prices gives by a lower bound for welfare losses, while that based on the old, observed prices provides an upper bound.

---

[5] Of course the reader also can compute ΔP_K from the tables

As capital services are fixed in aggregate supply by assumption, any flat-rate tax on capital income is nondistortionary. Further, since the tax proceeds are distributed according to one's ownership of capital, the distribution of income is unaffected by flat-rate capital income taxes. It is for these reasons that the only price which changes in moving from a flat 45.16878 percent tax (the surtax-removal cases) to a flat tax at an equal-yield rate is the net price of capital. For each of the seven production parameterizations, the new net price of capital and the equal-yield tax rate applying to net capital income are shown in the "P_K" and "tax rate" rows, respectively.

Part B of table 4 presents the same seven factor-elasticity cases where the consumers' elasticities of substitution are now specified as 0.5. As is shown, the efficiency loss estimates are reduced by 10–40 percent, and the shift factors are substantially lower, indicating that labor bears a smaller share of the burden of the surtax. The sensitivity of this incidence of the distortion on the functional distribution of income to different parameterizations seems to be far greater than expected or indicated by Harberger.

In comparing the estimates of the two parts of table 4 with Harberger's corrected results, one arrives at the general conclusion that the magnitudes are similar. The capital reallocation caused by the discriminatory capital income taxation is estimated to be smaller with the algorithmic approach of this section. Recalling the inequality of the units of measurement, the figures can best be interpreted by comparing the percentage increase in the amount of capital allocated to sector X which would occur in moving from the distortionary to a neutral tax situation. The six comparable

---

[6] What is referred to here as NNP is the value of the product of the two (or 12) sectors. The efficiency loss is the value of the change in production.

TABLE 4

SUMMARY OF TWO-SECTOR RESULTS

Case	Observed Equilibrium *	1	2	3	4	5	6	7
		A Consumer Demand Elasticities = 1, Fixed Factor Supplies						
S_X		−1 0	−1 0	−1 0	−0 75	−0 5	−0 5	−0.25
S_Y		−1 0	−0 5	−0 25	−0 25	−0 5	−0.25	−0.25
Surtax removal.								
P_X	1 00000	0 96626	0 97136	0 97419	0 96834	0 95607	0 95921	0.94337
P_Y	1 00000	1 17068	1 19455	1 20842	1 18378	1 13088	1 14478	1 07529
P_L	1 00000	1 00000	1 00000	1 00000	1 00000	1 00000	1 00000	1.00000
P_K	1 00000	1 29695	1 33027	1 34900	1 30727	1 22027	1 24145	1.12495
R	37 168	24 665	25 299	25 655	24 861	23 207	23 610	21.394
X^*	252 265	261 074	261.495	261 741	261.070	259 666	259 971	257 884
Y^* .	44.349	37 883	37 374	37 082	37 541	38 614	38 310	39.624
K_X^* ..	23.589	27 828	27 318	27 043	26 815	26.585	26 431	25.665
K_Y^*	18 515	14 276	14 786	15 061	15 289	15.519	15 673	16.439
L_X^* .	199 871	199 871	201 250	202 026	201 916	201 165	201 731	201.366
L_Y^*	17 471	17 471	16 092	15 316	15 426	16 177	15 611	15 976
Shift factor§		0 00000	−0 17158	−0 27634	−0 05125	0 33570	0 25018	0.66648
Relative share of rich	0 273	0 273	0 274	0 275	0 274	0.271	0 272	0.269
ΔNNP† .		0 943	0 633	0 450	0 467	0 591	0 477	0.435
ΔNNP‡ .		2 344	2 255	2 209	1 997	1 666	1.666	1 094
Equal yield replacement								
P_X	1 00000	0 96626	0 97136	0 97419	0 96834	0 95607	0 95921	0 94337
P_Y	1 00000	1.17068	1 19455	1 20842	1 18378	1.13088	1.14478	1 07529
P_L	1 00000	1 00000	1 00000	1 00000	1.00000	1 00000	1.00000	1.00000
P_K	1 00000	1 00692	1 04901	1 07269	1 01902	0 90749	0 93441	0 78472
R	37 168	36 877	37 141	37 289	36 998	36 376	36.538	35 719
Tax rate	1 22112/0 45169	0 86983	0 84091	0 82795	0 86232	0 95204	0 92871	1 08108

TABLE 4 (*Continued*)

		B Consumer Demand Elasticities = 0 5, Fixed Factor Supplies						
S_X		−1 0	−1 0	−1 0	−0 75	−0 5	−0 5	−0 25
S_Y		−1 0	−0 5	−0 25	−0.25	−0 5	−0 25	−0 25
Surtax removal:								
P_X	1 00000	0 97166	0 97842	0 98230	0 97781	0 96531	0 97032	0 95551
P_Y	1 00000	1 18989	1 22155	1 24078	1 22182	1 16734	1 18994	1 12585
P_L	1 00000	1 00000	1 00000	1 00000	1 00000	1 00000	1 00000	1 00000
P_K	1 00000	1 33226	1 37745	1 40398	1 37175	1 28306	1 31769	1 20961
R	37 168	25 337	26 196	26 701	26 088	24 401	25 060	23 004
X*	252 265	257 426	257 618	257 740	257 460	256 760	256 979	255 969
Y*	44 349	40 879	40 502	40 279	40 462	41 044	40 783	41 470
K_X*	23 589	26 862	26 180	25 800	25 693	25 760	25 506	25 096
K_Y*	18 515	15 242	15 924	16 304	16 411	16 344	16 598	17 007
L_X*	199 871	198 180	199 707	200 595	200 583	199 872	200 562	200 511
L_Y*	17 471	19 162	17 635	16 747	16 759	17 470	16 780	16 830
Shift factor		−0 18240	−0 44843	−0 62499	−0 41262	0 06652	−0 10472	0 37689
Relative share of rich	0 273	0 274	0 276	0 276	0 275	0 273	0 274	0 271
ΔNNP†		0 885	0 539	0 328	0 331	0 481	0 330	0 305
ΔNNP‡		1 690	1 507	1 405	1 308	1 190	1 148	0 831
Equal yield replacement								
P_X	1 00000	0 97166	0 97842	0 98230	0 97781	0 96531	0 97032	0 95551
P_Y	1 00000	1 18989	1 22155	1 24078	1 22182	1 16734	1 18994	1 12585
P_L	1 00000	1 00000	1 00000	1 00000	1 00000	1 00000	1 00000	1 00000
P_K	1 00000	1 04987	1 10675	1 14016	1 09892	0 98588	1 02976	0 89152
R	37 168	37 039	37 594	37 809	37 575	36 913	37 183	36 397
Tax rate	1 221122/0 45169	0 84215	0 80675	0 78759	0 81211	0 88928	0 85760	0 96964

* Billions of units

† $ billion Calculated at new (i e nondistortionary) prices

‡ $ billion Calculated at old (i e observed) prices

TABLE 5

% INCREASE IN CAPITAL ALLOCATED TO "CORPORATE" SECTOR WITH NEUTRAL CAPITAL INCOME TAXATION

S_X	S_Y	V	Estimate of Harberger Model (%)	Scarf Algorithm Estimate (%)
−1 0	−1 0	−1 0	19 74	17 97
−1 0	−0 5	−1 0	17 08	15 81
−0 5	−0 5	−1 0	13 87	12 70
−1 0	−1 0	−0 5	14 98	13 88
−1 0	−0 5	−0 5	11 68	10 98
−0 5	−0 5	−0 5	9 87	9 20

cases are shown in table 5 [7] The algorithmic estimates indicate that the discriminatory taxes cause slightly less resource misallocation and therefore a slightly smaller dead-weight loss These facts are confirmed by noting that table 4 also indicates a smaller change in the price of capital services than do the comparable cases in column 10 of table 3 In all cases the percentage of deviation between the two approaches correctly implemented is small, indicating that, despite its several shortcomings, Harberger's model does not give a poor approximation, even for analyzing tax distortions as large as these

Somewhat more ambiguity occurs in comparing the dead-weight loss estimates, although again the answers do not conflict I would argue, however, that, in giving a point estimate, the Harberger approach may give the user of that technique a false sense of precision While Harberger qualifies his results as rough estimates, the approach of this section gives an explicit upper and lower bound for the welfare loss [8] In the six comparable cases, the Harberger corrected point estimate is always well within the upper and lower bounds prescribed in table 4 and is somewhat larger than their arithmetic or geometric mean The geometric mean, of course, forms the Fisher "ideal" index (Fisher 1927) Again, however, the point estimate provided by the Harberger method does not deviate far from what one would arrive at with the technique of this section

[7] The two models are not strictly comparable One difference is the fact that two consumer classes are dealt with in this section, while the Harberger approach encompasses only a single-consumer or market demand relationship The Scarf algorithm method was also run with a single consumer, and the results in terms of resource allocation across the two production sectors, prices, and dead-weight loss were nearly identical with the examples with two consumer classes reported here So this does not account for the smal differences in the results between Harberger's method and those of this section The differences are due to the local assumptions of the Harberger model and the lack of a income effect in its demand specification

[8] Of course, for a multitude of reasons, the numerical accuracy of these bounds themselves needs qualification

The Scarf algorithm approach of this section allows the incorporation of a labor-leisure choice in the model A fairly simple and possibly unrealistic demand function for leisure was analyzed in Shoven (1973) and Shoven and Whalley (1972) The more recent study covered the same elasticity cases reported on here The general results were that the welfare loss estimates were approximately the same as in the fixed-labor-supply case, as long as leisure was valued at the wage rate It is possible that the value of production can actually be higher in the presence of the distortion than in its absence That is, the inefficiency of the distortion can be manifested totally in a decreased consumption of leisure The lesson from that earlier analysis, then, is that the change in NNP is likely to be a poor estimate of the efficiency loss if there is a substantial elasticity in the supply of labor

Given that the two approaches, when applied correctly to two production-sector data, are in broad agreement as to the incidence and efficiency effects of capital income taxes, the major advantage of the Scarf algorithmic method is its capability to analyze a more disaggregated description of production The Harberger producer-surplus approach can likewise be disaggregated, but the general equilibrium model he uses to solve for the capital reallocation cannot easily incorporate more than a two-way division of production, except for the case in which all elasticities of substitution are the same (Hoffman 1972)

In order to evaluate the impact of disaggregation, the 12-sector data of table 2 were analyzed using the Scarf algorithm The results for a selective subset of the 12 production-sector cases are contained in table 6 As intermediate products are not explicitly considered, it is necessary to interpret consumers as demanding value added from each of the 12 sectors. What is compared is the current tax situation (with each of the 12 sectors facing the different effective tax rates, as shown in column 9, table 2) with a flat-tax regime, with the tax rate set at 45 16878 percent (for comparability with the two-sector results just presented) As above, the equal-yield tax rate and the corresponding net price of capital were computed and are reported Compared with the two-sector results shown in the previous set of tables, the loss estimates for these 12-sector cases are 10–70 percent higher. The efficiency cost as a percentage of the surtax yield now goes as high as 24 4 percent The restrictiveness of the two-sector bifurcation of production is seen by noting that large relative price differentials develop between sectors which had previously been aggregated together in the "corporate" sector In particular, the price of petroleum and coal products is shown to rise as much as 24 percent relative to the price of communication and public utilities with the switch to neutral capital income taxation. While the two-sector results had indicated that the price of the "corporate" sector's output would fall, it is seen in table 6 that the price of some components of that sector would rise relative to the

TABLE 6

SELECTIVE SUMMARY OF 12 SECTOR RESULTS

Case*	Observed Equilibrium	1	2	3
	A Consumer Demand Elasticities = 1, Fixed Factor Supplies			
$S_{1,2,3}$		−1.0	−0 75	−0 25
S_{4-12}		−1 0	−0 25	−0 25
Surtax removal·				
P_1/Y_1	1.00000/ 16 281	1 13667/ 14 322	1 14850/ 14 197	1 07483/ 14 680
P_2/Y_2	1 00000/ 25 298	1 19444/ 21.180	1 20492/ 21 027	1.09027/ 22 487
P_3/Y_3	1 00000/ 2 770	1 15212/ 2 404	1 17168/ 2 368	1 10970/ 2 419
P_4/X_4	1 00000/ 3.216	0 96275/ 3 340	0 96454/ 3 339	0.94226/ 3.304
P_5/X_5	1.00000/ 17 865	0 99713/ 17 916	0 99753/ 17 937	0 98936/ 17 480
P_6/X_6	1 00000/104 291	0 93235/111 859	0 93539/111 670	0.91523/110.309
P_7/X_7	1 00000/ 3 144	1 01668/ 3 092	1 01817/ 3 093	0 98789/ 3 081
P_8/X_8	1 00000/ 4 874	1 07470/ 4.535	1 07906/ 4 524	0.99339/ 4.750
P_9/X_9	1.00000/ 54 487	1 00463/ 54 296	1.00582/ 54 257	0 98013/ 53 816
P_{10}/X_{10}	1 00000/ 16 761	0 96786/ 17 318	0.96940/ 17 317	0 95340/ 17 019
P_{11}/X_{11}	1 00000/ 13 883	0 87008/ 15 956	0 87477/ 15 895	0 83382/ 16 118
P_{12}/X_{12}	1 00000/ 33 744	0 99968/ 33 755	1 00003/ 33 796	0 99244/ 32 915
P_L	1.00000	1 00000	1 00000	1.00000
P_K	1 00000	1 29695	1 30453	1 14267
R	37 168	24 665	24 809	21 731
Shift factor		0 00000	−0 03749	0 610853
ΔNNP†		1 344	0 774	0 477
ΔNNP‡		3 300	2 808	1 762
Equal yield				
P_K	1 00000	1 00971	1 01792	0 80915
R	37 168	36 759	36 877	35 774
Tax Rate		0 86466	0 86044	1 05006
	B Consumer Demand Elasticities = 0 5, Fixed Factor Supplies			
$S_{1,2,3}$		−1 0	−0 75	−0 25
S_{4-12}		−1 0	−0 25	−0 25
Surtax removal				
P_1/Y_1	1 00000/ 16.281	1 14982/ 15 276	1 17555/ 15 144	1.10548/ 15.371
P_2/Y_2	1.00000/ 25 298	1 21625/ 23 079	1 24727/ 22 845	1 13786/ 23 542
P_3/Y_3	1.00000/ 2 770	1 16216/ 2 585	1 19441/ 2 556	1 13551/ 2 580
P_4/X_4	1.00000/ 3 216	0 96788/ 3 290	0 97361/ 3 289	0 95204/ 3 270
P_5/X_5	1 00000/ 17 865	0 99879/ 17.993	1 00053/ 18 025	0.99278/ 17 790
P_6/X_6	1 00000/104 291	0 93784/108 395	0 94490/108 277	0 92506/107 588
P_7/X_7	1 00000/ 3 144	1 02246/ 3.130	1.02890/ 3 128	1.00064/ 3 119
P_8/X_8	1 00000/ 4 874	1 09141/ 4 696	1 11011/ 4 669	1 02938/ 4 767
P_9/X_9	1 00000/ 54 487	1 00963/ 54 580	1 01501/ 54 581	0 99086/ 54 311
P_{10}/X_{10}	1 00000/ 16 761	0 97172/ 17 114	0 97616/ 17 121	0 96059/ 16 968
P_{11}/X_{11}	1 00000/ 13 883	0 88024/ 14 894	0 89270/ 14 829	0 85279/ 14 916
P_{12}/X_{12}	1 00000/ 33 744	1 00118/ 33 944	1 00276/ 34 008	0.99561/ 33 555
P_L	1.00000	1.00000	1.00000	1 00000
P_K	1 00000	1 32957	1 36463	1 20969
R	37 168	25.286	25 952	23 006
Shift factor		−0 16777	−0 36883	0 37659
ΔNNP†		1.298	0 657	0 390
ΔNNP‡		2 363	1.857	1 164
Equal yield				
P_K	1.00000	1.04844	1.09151	0.89259
R	37.168	37 122	37.452	36.357
Tax rate		0.84094	0.81494	0.96741

* Sector 1, agriculture, 2, real estate, 3, crude oil and gas, 4, mining; 5, contract construction, 6, manufacturing, 7, lumber and wood products, 8, petroleum and coal products; 9, trade; 10, transportation, 11, communication and public utilities, 12, services

† $ billion. Calculated at new (i.e , nondistortionary) prices

‡ $ billion. Calculated at old (i e , observed) prices

TABLE 7

FISHER INDEX OF SOCIAL WASTE
CONSUMER DEMAND ELASTICITIES = 1.0, FIXED FACTOR SUPPLIES

2 SECTORS			12 SECTORS		
S_X	S_Y	$ Billion	$S_{1,2,3}$	S_{4-12}	$ Billion
−1.00	−1.00	1.49	−1.00	−1.00	2.11
−1.00	−0.50	1.19			
−1.00	−0.25	1.00			
−0.75	−0.25	0.97	−0.75	−0.25	1.47
−0.50	−0.50	0.99			
−0.50	−0.25	0.89			
−0.25	−0.25	0.69	−0.25	−0.25	0.92

price of labor. The added detail of this disaggregated analysis thus gives a somewhat different picture and information which may be valuable in the making of policy decisions.

As has been shown in tables 4 and 6, a change in tax rates causes a vector of changes in outputs which must be aggregated to arrive at a single-number dead-weight loss estimate. This condensation of information is subject to the usual index number problems, and I have presented the Paasche and Laspeyres measures of the inefficiency resulting from the distortionary taxation of capital income in the 1953–59 U.S. economy. Nonetheless, policy makers often want a single number, and not a range of numbers (they want the impossible solution to the index number problem). One technique of combining the Paasche and Laspeyres indices into a single loss estimate is to form the Fisher "ideal" index (Fisher 1927), which is the square root of their product. The results of such a procedure for the cases shown in part *A* of tables 4 and 6 are shown in table 7. If personally asked to give a point estimate of the efficiency loss due to the U.S. capital income taxes, I would refer to the $S_X = -0.75$, $S_Y = -0.25$ 12-sector model. The Fisher index shows a loss of $1.47 billion in this case, or 11.9 percent of the surtax revenue. The point estimate is somewhat smaller than what one would arrive at from Harberger's published estimates, although, given the magnitude of his two mistakes and the fact that this is for a 12-sector rather than a two-sector analysis, the number is remarkably close to the $2.0–$2.5 billion figure one might get from his work. Needless to say, one would still want to qualify heavily any such point estimate.

IV. Conclusion

This paper has reexamined the results of the Harberger model concerning the static efficiency cost of the U.S. taxation of capital income during the 1953–59 period. I have asserted that Harberger made two significant mistakes in his earlier evaluation of this question, each of which substantially altered his results. The net effect of correcting these errors is to lower the dead-weight loss estimates previously published as ranging from $1.0 billion to $2.9 billion approximately 38 percent to a range of $0.625 billion–$1.79 billion. The correct results are thus an even smaller fraction of GNP or NNP than previously thought, but the efficiency loss does amount to between 6 and 15 percent of the revenue generated by the distortionary surtax on capital income originating from the "corporate" sector.

Using the corrected data, I have computed the incidence and dead-weight loss using an algorithmic-solution technique for a general equilibrium model. In the two-sector analyses, the results do not differ substantially from the corrected Harberger figures, showing, if anything, that the unequal tax treatment causes less resource reallocation. The conclusion is that, at least in this case, the second-order approximations made in the Harberger analysis are reasonably accurate, even for tax distortions as large as those of this study. The algorithmic approach does explicitly remind one of the fundamental index number problems in describing the static inefficiency effects of a tax distortion with a single number.

One of the advantages of the algorithmic method is its capability of solving a disaggregated general equilibrium model. In order to evaluate the effect of moving to a less aggregated description of production, a 12-sector analysis was presented. The dead-weight loss estimates increased an average of approximately 40 percent from what they were in the two-sector cases. The model also implied that differential capital income taxes caused significant relative output price distortions among sectors which were previously lumped together. It was stated that the added richness of the results of the disaggregated analysis should be valuable to those involved in tax policy evaluation and formation.

The final point which should be made is that, in comparing the Harberger model and approach with the algorithmic-solution procedure in this study, I have not pressed the latter method anywhere near its capabilities. Because of its ability to handle models with many commodities and consumers, detailed incidence studies are feasible. Further, the additional disaggregation may allow studying several distortions simultaneously or, by time-dating commodities, permit a dynamic evaluation of policy alterations. The purpose here was to correct Harberger's errors in his study of the efficiency cost of uneven capital income taxation and to examine how different the results of a similar simple analysis would be with this algorithmic approach. The implementation of the further capabilities of the algorithmic technique in conjunction with more recent economic data is an area of active research for this author and others but is beyond the scope of this study.

References

Fisher, I. *The Making of Index Numbers: A Study of their Varieties, Tests, and Reliability.* 3d ed. New York: Houghton Mifflin, 1927.

Harberger, A. C. "The Corporation Income Tax An Empirical Appraisal " In *Tax Revision Compendium* Vol I Washington House Committee on Ways and Means, Government Printing Office, 1959

———. "The Incidence of the Corporation Income Tax " *J P E* 70, no 3 (June 1962) 215–40

———. "Efficiency Effects of Taxes on Income from Capital " In *Effects of Corporation Income Tax*, edited by M Krzyzaniak Detroit· Wayne State Univ Press, 1966

Hoffman, R F "Disaggregation and Calculations of the Welfare Cost of a Tax." *J P.E* 80, no 2 (March/April 1972) 409–17.

Lucas, R E "Labor-Capital Substitution in U S Manufacturing " In *The Taxation of Income from Capital*, edited by A C Harberger and M. J Bailey Washington Brookings Inst , 1969

McLure, C E , Jr "A Diagrammatic Exposition of the Harberger Model with One Immobile Factor " *J P E* 82, no 1 (January/February 1974) 56–82

——— "General Equilibrium Incidence Analysis The Harberger Model after Ten Years " *J Public Econ* 4 (February 1975) 125–61

Minasian, J. R "Elasticities of Substitution and Constant-Output Demand Curves for Labor " *J P E* 69, no 3 (June 1961) 261–70

Projector, D S , and Weiss, S *Survey of Financial Characteristics of Consumers* Federal Reserve Technical Papers Washington Government Printing Office, 1966

Rosenberg, L. C "Taxation of Income from Capital, by Industry Group " In *The Taxation of Income from Capital*, edited by A C Harberger and M J Bailey. Washington Brookings Inst , 1969

Scarf, H E , with the collaboration of Hansen, T *The Computation of Economic Equilibria*. New Haven, Conn Yale Univ Press, 1973

Shoven, J B "General Equilibrium with Taxes Existence, Computation, and a Capital Income Taxation Application." Ph D dissertation, Yale Univ., 1973.

——— "A Proof of the Existence of a General Equilibrium with Ad Valorem Commodity Taxes " *J Econ Theory* 8 (May 1974) 1–25

———. "Applying Fixed Point Algorithms to the Analysis of Tax Policies " In *Fixed Points · Algorithms and Applications*, edited by C B Garcia and S Karamardian. New York · Academic Press, 1976

Shoven, J. B., and Whalley, J "A General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U S " *J Public Econ.* 1 (November 1972) 281–321

———. "General Equilibrium with Taxes A Computational Procedure and an Existence Proof " *Rev Econ Studies* 60 (October 1973) 475–90

———. "Equal Yield Tax Alternatives A General Equilibrium Computational Technique " Technical Report no 150(R), Stanford Univ , Inst. Math. Studies Soc. Sci., 1975

Solow, R. M. "Capital, Labor, and Income in Manufacturing " In *The Behavior of Income Shares*. Princeton, N.J. Princeton Univ. Press (for Nat. Bur. Econ. Res.), 1964.

Stiglitz, J. E. "The Corporation Tax " Technical *Report no 162, Stanford Univ.*, Inst. Math. Studies Soc. Sci., 1975.

Van Horne, J. C. *Financial Management and Policy*. Englewood Cliffs, N.J : Prentice-Hall, 1971.

Whalley, J. "A Numerical Assessment of the April 1973 Tax Changes in the United Kingdom." Ph.D. dissertation, Yale Univ., 1973.

Communications

The Incidence and Efficiency Effects of Taxes on Income from Capital: A Reply

Arnold C. Harberger

University of Chicago

Neil Bruce

Queen's University

I

Having attention called to one's mistakes cannot be classified as a happy event. Nonetheless, it must be done when mistakes of significance are involved, and I am glad that John Shoven (pp 1261–83, this issue) has minimized bloodshed in carrying out the task

I have no alternative but to plead guilty to both of the errors of which I am accused One of them was a straightforward arithmetic mistake, for which I alone am responsible [1] The second was a blunder in applying a theoretical concept that is basically crystal clear Its application, however, sometimes requires a certain degree of subtlety, which I shall attempt here briefly to describe

The subtlety has to do with the circumstances under which, and the senses in which, the welfare costs of a tax system are independent of the

---

This reply represents a rather special form of joint authorship Section I, acknowledging errors in his earlier paper, is exclusively Harberger's Section II, discussing the accuracy of alternative procedures for estimating welfare costs, is joint in the usual sense And the Appendix, setting out the calculations of compensating and equivalent variations, is exclusively Bruce's, whose idea it was to use these as the standard of comparison in Section II.

[1] No research assistant helped me with the calculations in question, hence I have no one with whom to share the blame The lessons I have learned from this particular error are (*a*) do not write papers in marathon sessions of more than 24 consecutive hours; (*b*) always, but especially so if rule *a* is violated, have the key calculations checked by someone else.

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved.

addition of a neutral tax. In a certain sense, this independence has to hold if the added tax is indeed neutral. Yet it is quite clear that the procedures I used for calculating welfare costs would have been correct had the distorting tax been the only tax in the system. The error came in my failing to see that, at least in *some* cases, my measure of the welfare cost of the stated distortion would be different in the presence of a neutral tax from the measure to be used in its absence.

To put this explanation on the simplest ground possible, consider the simple textbook case of a world with linear demand relationships and with only two commodities, A and B, produced at constant unit resource costs ($c_a = c_b = \$1$) under conditions of factor supply such that $Ac_a + Bc_b = R$, a constant. In exploring this case, let us consider three alternative tax packages: (1) a specific (distorting) tax of T_a cents per unit of A, combined with another specific (neutral) tax of T_n cents per unit of A and B produced (and consumed) in this stylized economy, (2) an ad valorem (distorting) tax of τ_a percent of the resource costs of A, combined with an ad valorem neutral tax of τ_n percent of the resource costs of both A and B, (3) an ad valorem (distorting) tax of τ_a^* percent of the price of A, that price being defined so as to include the neutral tax τ_n applying to the resource costs of both A and B.

In case 1, the standard welfare-cost triangle for the distorting tax is simply $\frac{1}{2}T_a|\Delta A|$, this measure does not change as T_n is introduced or changed. In case 2, the welfare cost of the distorting tax is $\frac{1}{2}\tau_a c_a|\Delta A|$, this too is invariant with respect to the introduction of or changes in τ_n. In case 3, however, the welfare cost is $\frac{1}{2}\tau_a^*(1 + \tau_n)c_a|\Delta A|$, this is obviously not independent of τ_n.

In the set of calculations that Shoven is criticizing, I distilled the taxation of income from capital in the United States into a neutral tax (like τ_n) of 45 percent and a distorting tax of 85 percent. This tax was like τ_a^* in case 3 above, in that the neutral tax and the distorting tax combined in a multiplicative fashion to link resource cost with market price (i.e., $c_a[1 + \tau_n][1 + \tau_a^*] = P_a$). Yet, in interpreting and working with the figure, I acted as if it were like τ_a in case 2 (i.e., $c_a[1 + \tau_n + \tau_a] = P_a$). Shoven's point, with which I fully agree, is that, to be consistent with this interpretation, my measure of the distortion should have been 1.45 × 85 percent, or 123 percent. Alternatively and equivalently, the base to which the 85 percent distortion was applied should itself have been augmented by 45 percent.

Readers may not even yet appreciate the degree of subtlety that is involved here. As I have expounded the cases, it seems that T_a in case 1 (or τ_a in case 2) gives a measure of welfare cost that is invariant to introducing or changing T_n (or τ_n in case 2), while τ_a^* in case 3 gives a measure that is not invariant. All these conclusions are true, providing that one is taking unit resource cost as the numeraire in terms of which one measures welfare cost.

But it is not at all clear that unit resource cost is the most convenient or appropriate numeraire. If we use the price paid by consumers for the product of the undistorted sector (P_b) as the numeraire, we find that our measure of welfare cost in cases 1 and 2 ends up varying with T_n and τ_n, while our measure in case 3 ends up invariant to changes in τ_n. The same conclusion applies if total consumer expenditures is taken as the numeraire—the choice that seems to have the best theoretical justification.[2]

These results jibe well with common sense. A tax of \$0.10 per unit of a product whose resource costs are \$1.00 per unit will, for linear demand curves, have the same dollar value of welfare costs when there is also a neutral tax equal to 50 percent of cost as when there is no neutral tax at all. But the value of those welfare costs, measured in terms of consumer expenditures, will be lower (I am assuming here that all situations are comparable in the sense that $c_a = c_b = \$1.00$). In this way the *real* costs of T_a and of τ_a are lower in the presence of a neutral tax T_n or τ_n than in its absence. But, if the distorting tax is expressed as a fraction of the price the consumer pays (either gross or net [as in τ_a^*] of the distorting tax itself), then the *real* situation faced by the consumer is invariant to the introduction of or changes in τ_n, even though the nominal dollar measure of the welfare cost of the tax changes as a function of τ_n. And this invariance of the consumer's real situation is picked up when the numeraire is either total consumer expenditure or the price (P_b) of a good subject only to neutral taxation. It is missed when the numeraire is unit resource cost.

If this discussion serves to convince readers that some quite tricky issues are involved here, one of my purposes will already have been served. And, if readers conclude from the discussion above that the case 3 way of providing a quick and stylized snapshot of a tax system has advantages over those of cases 1 and 2, in the sense that the invariance involved is "real" in a deeper and different sense, then my second purpose will have been served. My mistake was not in choosing to use the case 3 model but in failing to carry through the calculations in a fashion consistent with that choice.

II

We now turn to the parts of Shoven's paper in which he advances and espouses an alternative solution to the problem in question—the algorithmic approach. When all is said and done, this approach amounts to setting out a full (even though highly simplified) general-equilibrium model, specifying production and demand functions, resource constraints, and other relevant parameters. Alternative solutions of this model are then found for the "with tax" and "without tax" cases.

---

[2] Diewert (1966) shows that what he calls the Harberger welfare indicator (a generalized consumer-surplus triangle) is exact for quadratic utility functions when normalized on total expenditure. Bruce (1975, 1977) extends this result to all homothetic utility functions, for a generalized triangle whose hypotenuse can be curvilinear (where Diewert's are linear).

At this level of operations, we have no quarrel with the algorithmic approach It certainly is not a bad thing to see the full range of the functions one is dealing with and to calculate explicitly the full solutions that the system yields under alternative sets of assumptions

Our only debate concerns what is done with the solutions once they are obtained In the published applications to date of the algorithmic approach (Shoven and Whalley 1972, 1973, Shoven 1973, 1974, 1976), the typical procedure has been to obtain the vector of changes ΔX_i in the quantities of the individual final products involved in the analysis and to "price" these changes at the prices prevailing either in the "without tax" or in the "with tax" equilibrium This procedure has the defect of implicitly attributing a rectangle of welfare cost, in a simple excise-tax example like that set out in the preceding section, when "with tax" prices are used; alternatively, it would assign a zero welfare cost to an excise tax when "without tax" prices (c_i) are used, as long as the production constraint is linear (i e , $\sum X_i c_i = R$) [3]

Perhaps partly as a result of our communicating these reservations informally, but probably more importantly as a result of his further independent consideration of the issues, Shoven in the present study adopts the convention of taking the geometric mean of the welfare costs measured at prices "without" and "with" the tax distortion The procedure, which is similar to valuing the ΔX_i using a Fisher ideal price index, yields a triangle of welfare cost in the simple textbook case (although with a curvilinear hypotenuse rather than the linear one that would emerge if the arithmetic mean of the two prices had been used). We welcome this modification, which makes it quite clear that our earlier objection did not call into question the algorithmic approach as such but only the particular way in which it was "married" to applied welfare economics in the exercises under discussion.

Our questions now are, How well does this modified approach do? How do the results compare with attempts to estimate consumer-surplus triangles directly? And, finally, are there ways in which the procedure can be further refined and improved?

Shoven's paper provides us with a useful way to answer the first question, in that one can specify a utility function to correspond to each of the demand functions used in the algorithmic approach We will deal, for simplicity, with his case 1, in which all three elasticities of substitution are equal to unity; but the methodology that we employ (see Appendix) can readily be extended to the other cases as well Once we have a utility function, we can proceed to measure directly the compensating variation in income (i.e , that income change, at the distorted prices, which would be necessary to bring consumers back up to their undistorted level of utility). And we can also measure the equivalent variation (i e , the amount by which income at undistorted prices could be reduced, while still permitting consumers to attain the same level of utility as they enjoy in the distorted situation) For case 1, the compensating variation turns out to be $1 675 billion, and the equivalent variation is $1 665 billion (see Appendix) Note that both the compensating and the equivalent variation are *exact* and unambiguous measures of the change in welfare They differ only with respect to the set of relative prices at which the welfare levels associated with two specific indifference curves are quantified and compared They are both therefore "right" answers to the first question. And it is worthwhile pointing out that comparing the same two welfare levels at some intermediate set of relative prices would, in general for well-behaved functions, give a result somewhere between these two bounds. Since in any case the bounds are so close together, we shall work from here on with $1 67 billion as a summary figure for "the truth."

---

[3] The results are more complicated with curvilinear production constraints or with factor-market distortions.

This figure can be compared with Shoven's own answers of $0 943 billion at undistorted prices and $2 34 billion at distorted prices (his table 4, pt A, col 1 rows 17 and 18) or their geometric mean of $1 49 billion (his table 7, col 3, row 1) The original Paasche and Laspeyres indexes generate incredibly wide margins, while the geometric mean is a quite close approximation to "the truth " (Note that we are not testing here the algorithmic solution, as that is used for obtaining all four figures [0 943, 1 49, 1 67, and 2 34 billion dollars], we are rather, checking the accuracy of the Paasche, Laspeyres, and Fisher [geometric mean] indexes of welfare change against our $1 67 billion figure for "the truth ")

The corresponding value for Harberger's estimate, as corrected by Shoven for both of the mistakes that motivated his article, is $1 79 billion This is a bit closer to "the truth" than Shoven's geometric mean estimate, even though it does not use the data generated by the algorithmic approach at all [4] But note that a straightforward arithmetic mean, using the algorithmic data, would come closer yet—\$1.644 billion—to the \$1.67 billion mark.

---

[4] The closeness of the corrected Harberger estimate to "the truth" seems to be a fully adequate demonstration that the several alleged "shortcomings" of Harberger's approach, associated with local or "small change" assumptions, are not as severe as Shoven's text implies Shoven himself does not come as close in this particular case Even though he calculates quantity changes exactly using the algorithmic method, his methods of valuing the social cost of these changes are sufficiently crude to offset the advantage of his better quantity-change figures. The comment above covers Harberger's use of linear equations as such, as well as whatever "inconsistency" may be involved in applying linear assumptions to large changes. We take this opportunity also to comment on another point that Shoven has made, to the effect that, for Harberger's procedure to be valid, "the labor allocation must be assumed to be unaltered, or the marginal product of capital must be taken as independent of the amount of labor employed in each sector Without either of

It is not our purpose, however, to quibble about details. In particular, we are confident that there will be times when a geometric mean of prices will prove more accurate than an arithmetic mean. Our conclusion from the exercise above is simply that all three figures, \$1.49 billion (Shoven's geometric mean, using algorithmic ΔX_i), \$1.644 billion (arithmetic mean, using algorithmic ΔX_i), and \$1.79 billion (Harberger's method, corrected for both mistakes) come acceptably close to "the truth" —far more so than either the \$0.943 billion or the \$2.34 billion figures generated by using one end-point set of prices or the other.

We reiterate that we have no quarrel with the algorithmic approach as such. It simply says, "Find the ΔX_i by specifying and solving a model under alternative assumptions about its parameters." The only question is whether it is worth while to do so if one is in considerable ignorance about the basic parameters as such. For example, if it takes a lot of work (or high-priced computer time) to calculate the algorithmic solution, its greater accuracy in drawing the exact ΔX_i implied by a specific set of parameters might not be much of a virtue in a situation where we are quite uncertain as to whether the key elasticities are -0.5, -1.0, or -1.5. Under such circumstances, the most economical way to proceed may be to follow Harberger, in acting as if one can use for discrete changes a formula that applies strictly only to differentials and accepting the (often modest) errors that this pretense might entail.

The truly elegant solution to the problem would be a Divisia index of welfare cost, obtained by doing sequential algorithmic solutions in which, say, the vector of taxes (or other distortions being analyzed) would be moved from zero to their final figure in successive small steps. If this calculation were performed (following Shoven) on changes in final-product quantities, it would entail applying to the successive sets of ΔX_i the average of prices at the beginning and end of the step. For small steps,

---

these conditions, the marginal product curves themselves will not be stationary" (p. 1264). Our response here is, "Perhaps true, in a sense, but surely irrelevant." It is true that, for the sectoral marginal product of capital to be treated as a function of the sectoral endowment of capital *alone*, the conditions above apply. However, it is not in any sense required, for the validity of the *welfare loss measure* $-\frac{1}{2}T(\Delta K_x)$, that the marginal product of capital be a function of K_x alone. Here one is dealing in principle with a reduced form, with T as the policy variable. The general expression for the welfare cost of a single distorting tax T^* on capital in x is

$$\int_{T=0}^{T^*} T(\partial K_x/\partial T)\, dT$$

The triangle formulation treats $(\partial K_x/\partial T)$ as constant over the relevant range, but that is the only assumption required. There is no need, in particular, for the marginal product of capital to be independent of the amount of labor or for the amount of labor in the affected activity to be constant. All this is treated in more detail in Harberger (1964).

it would not matter whether one used a geometric or an arithmetic mean Indeed, for small enough steps, one could even use "beginning of step" or "end of step" values and not worry about taking an average In doing so one would be approximating, as closely as one wished, the welfare-cost triangles of textbook examples (regardless of whether they are linear or curvilinear) and the line integrals that Hotelling (1938), Harberger (1964, 1971), and others have used to represent the welfare costs of a set of distortions in a general equilibrium context

Appendix

Calculation of Compensations and Equivalent Variations.

In the two-good case, the demand function

$$X = \frac{Ia_x}{p_x^v(a_x p_x^{1-v} + a_y p_y^{1-v})} \tag{A1}$$

derives from the CES utility function

$$u = A(\alpha_x X^{-\rho} + \alpha_y Y^{-\rho})^{-1/\rho}, \tag{A2}$$

where $v = -[1/(\rho + 1)]$ and $a_i = x_i^v/\alpha_y^{v-1}$ for $i = X, Y$ When $\rho = 0$, equation (A2) reduces to the Cobb-Douglas form

$$u = AX^{\alpha_x}Y^{\alpha_y}, \tag{A2'}$$

where α_i is the proportion of income I spent on commodity i and $\alpha_x + \alpha_y = 1$

The compensating variation is found by calculating

$$\left(\frac{u}{A}\right)_{NT} = (261\,074)^{8505}(37\,883)^{1495}$$

$$= 195\,623$$

using Shoven's no-tax values of X and Y

The level of income necessary to give this level of utility when the consumer faces distorted prices ($P_X = P_Y = P_x/P_y = 1$) is found by finding X and Y such that $X^{8505}Y^{1495} = 195\,623$ and which satisfy the first-order conditions in the distorted case, namely,

$$\frac{u_X}{u_Y} = \frac{8505Y}{1495X} = \frac{P_X}{P_Y} = 1$$

The foregoing are solved for $X = 253\,689$ and $Y = 44\,599$ and the necessary level of income $I_N = (1)X + (1)Y = 298\,289$ Subtracting actual national income at distorted prices (296 614) from I_N, we get the *compensating variation* as $1.675 billion.

The equivalent variation is found by calculating

$$\left(\frac{u}{A}\right)_T = (252\,265)^{8505}(44\,349)^{1495}$$

$$= 194\,525$$

The level of income needed to obtain this level of utility when facing undistorted prices ($P_X = .96626$, $P_Y = 1.17068$, and $P_X/P_Y = .8254$) is found by calculating

X and Y such that $194.525 = X^{8505} Y^{1495}$ and

$$\frac{u_x}{u_y} = \frac{.8505Y}{.1495X} = \frac{P_X}{P_Y} = .8254,$$

where the latter is the first-order condition at undistorted prices. We find $X = 259.6082$ and $Y = 37\ 6705$ and can calculate

$$\begin{aligned} I_T &= .96626X + 1\ 17068Y \\ &= 294\ 949 \end{aligned}$$

Subtracting I_T from actual national income at undistorted equilibrium (296.614), we get the *equivalent variation* as \$1 665 billion.

References

Bruce, Neil "Topics on Taxation in a General Equilibrium Economy" Ph D dissertation, Univ Chicago, 1975.

——— "A Note on Consumer Surplus, the Divisia Index, and the Measurement of Welfare." *Econometrica* (1977), in press.

Diewert, W E "Harberger's Welfare Indicator and Revealed Preference Theory" *A.E.R.* 66 (March 1966). 143–52.

Harberger, Arnold C. "The Measurement of Waste." *A.E.R Proc.* 54 (May 1964) 58–76

———. "Three Basic Postulates for Applied Welfare Economics An Interpretive Essay." *J. Econ Literature* 9 (September 1971) · 785–97

Hotelling, H "Demand Functions with Limited Budgets" *Econometrica* 6 (March 1938): 242–69

Shoven, J. B. "General Equilibrium with Taxes Existence, Computation, and a Capital Income Taxation Application" Ph D dissertation, Yale Univ, 1973

———. "A Proof of the Existence of a General Equilibrium with ad Valorem Commodity Taxes." *J Econ Theory* 8 (May 1974). 1–25

———. "Applying Fixed Point Algorithms to the Analysis of Tax Policies" In *Fixed Points. Algorithms and Applications*, edited by C. B. Garcia and S. Karamardian New York Academic Press, 1976

Shoven, J B., and Whalley, J. "A General Equilibrium Calculation of the Effects of Differential Taxation of Income from Capital in the U S" *J Public Econ* 1 (November 1972): 281–321

——— "General Equilibrium with Taxes. A Computational Procedure and an Existence Proof." *Rev. Econ. Studies* 60 (October 1973): 475–90.

Optimum Replacement of Capital Goods with Labor-saving Technical Progress: A Comparison of the Early New England and British Textile Firm

Peter L. Swan

Australian National University

In a recent article in this *Journal,* Williamson (1971) tried to capture the essence of Habakkuk's (1962) characterization of Anglo-American differences in replacement and scrapping policy in the 1820s as a response to different wage rates and rates of labor-saving technical change Habakkuk (1962, p 59) argued that "in so far as technical progress took the form of inventing machines which saved labour, but with some increase (or at least no demonstrable saving) in capital cost it might pay the Americans to scrap when it did not pay the English " The greater readiness to scrap in the United States was attributed by Habakkuk to the higher relative wage in America compared with Great Britain, and this proposition formed the core of his argument Thus it is surprising that Williamson (1971, p 1327) should conclude that, in his "model of labor saving (based on Habakkuk's hypothesis), unskilled wages do not appear to influence the replacement decision at all'" In fact, Williamson went further, to claim that "it appears that Habakkuk's model is not internally consistent " The present paper will demonstrate that Habakkuk was right in his conjecture that higher wages would induce more rapid replacement and that it is Williamson's, not Habakkuk's, model which is internally inconsistent.

Denslow and Schulze (1974) have pointed out two implicit assumptions made by Williamson (*a*) the net present value of a machine is independent of the date of its installation and hence of the best-practice technology of that date embodied in the machine, and (*b*) the economic machine life is constant despite the fact that the life must ultimately become infinite Both these assumptions are inconsistent with the other assumptions made

I wish to thank Edward Sieper for discussions which led to the writing of this paper Warren H. Hausman of the University of Rochester developed the initial dynamic programming approach to the problem which formed a basis for the solution ultimately adopted. Peter Hammond of the University of Essex made an important contribution in establishing the method of solution finally adopted in Sec I, i.e , eqq (1)–(9), and obtaining in addition the analytical solution for the special case Rick Bellas provided invaluable research assistance. Responsibility for any errors rests entirely with myself

[*Journal of Political Economy*, 1976, vol 84, no. 6]
© 1976 by The University of Chicago. All rights reserved

by Williamson of labor-saving technical advance which reduces the requirements of only one factor of production and a constant wage rate. Denslow and Schulze (1974) do not attempt to solve correctly Williamson's original problem but, instead, use the approach adopted in the growth literature on putty-clay models, in which the wage rate is assumed to rise by the amount necessary to offset the productivity rise for labor (see, e g , Bliss 1968). In Williamson's (1974) reply to Denslow and Schulze, he accepts their correction of misspecification (*a*) and completely ignores misspecification (*b*) but claims that his earlier conclusions remain basically unchanged (Williamson 1974, p 639).

The aim of the present paper goes beyond a vindication of Habakkuk's result, to develop a formal model of the optimum scrapping decision due to obsolescence when new labor-saving (Harrod-neutral) techniques must be embodied in new equipment and machinery The model developed in Section I is microeconomic and partial equilibrium in nature, insofar as it considers a single firm or industry subject to labor-saving technical advance facing an exogenously determined (and constant) wage rate and machine price. Thus the aim is to solve for the optimum stream of replacement decisions on the basis of Williamson's assumption of a constant wage and machine price and at the same time to relax the inconsistent assumption *b* of a constant machine life irrespective of installation date

It is shown that the age at replacement of a chain of machines is continually rising over time and that a rise in the wage rate or fall in machine price increases optimum lives An analytical solution to the problem is derived for the special case in which the rate of interest equals the rate of Harrod-neutral technical progress.

In Section II historical data is used to make quantitative predictions of optimum scrapping dates for American and British textile firms on the basis of the model Comparisons are made between the actual and predicted economic lives for the nineteenth-century New England textile industry Moreover, light is shed on the reasons for differences in Anglo-American firm behavior using 1830 data

A by-product of these investigations is that few of the theoretical or quantitative results obtained by Williamson (1971, 1974) survive close examination, and consequently virtually no support is found for his contention that it is "remarkable how little the conclusions of the earlier argument require revision" (Williamson 1974, p. 639).

I. A Model of Optimum Replacement with Labor-saving Technical Change

Suppose that it is time $t = 0$ and that a machine was last installed at time $t_0 < 0$. What is the optimum sequence of replacement times when

a) The cost of manning a machine is $\alpha = b(0)w$, where $b(0) = L(0)/K(0) =$ labor requirements per unit of pro ductive capacity in some initial period $t = 0$ (where L is the number of homogeneous labor inputs and K the number of identical machines) and w = the wage rate

b) The time-stationary price of a new unit of productive capacity is p, which is equivalent to the price of a single machine If there is technological change in the machine-producing industry, then p would be falling over time This could be allowed for, but, in keeping with Williamson's analysis, this complication will be neglected

c) The rate of Harrod-neutral (i e , labor-saving) technological change embodied in new machines is $\mu > 0$

d) The rate of interest is $\delta > 0$ (assumed constant)

In an intertemporal context, under the conditions specified, minimization of the present value of the cost of producing a unit of output in perpetuity represents a necessary condition for the maximization of the present value of net receipts (i.e , profits) Thus the criterion to be employed in selecting replacement dates, namely, the minimization of the present value of cost outlays, is appropriate regardless of the market structure in which the firm is operating [1]

Let $t_1, t_2,$, t_R be the times at which new machines are installed Then the cost of producing the constant output stream forever is

$$C = \int_0^{t_1} \alpha e^{-\mu t_0} e^{-\delta t}\, dt + p e^{-\delta t_1} + \int_{t_1}^{t_2} \alpha e^{-\mu t_1} e^{-\delta t}\, dt + p e^{-\delta t_2} + \quad + \quad + \int_{t_R}^{\infty} \alpha e^{-\mu t_R} e^{-\delta t}\, dt \tag{1}$$

$$= \frac{\alpha e^{-\mu t_0}}{\delta} + \sum_{r=1}^{R} \left\{ \frac{\alpha}{\delta} \left[e^{-(\mu+\delta)t_r} - e^{-\mu t_{r-1} - \delta t_r} \right] + p e^{-\delta t_r} \right\} \tag{2}$$

For simplicity we define the following new variables $\omega = p\delta/\alpha$ (the ratio of capital rental payments to wage payments , $x_r = e^{-\mu t_r}\omega$, $\varepsilon = \delta/\mu$, and $c = \omega^{-\varepsilon}\{(C/p) - [(\alpha e^{-\mu t_0})/(p\delta)]\}$

Because t_0 is fixed by history, it is sufficient to choose the variables $x_r (r = 1 \text{ to } R)$ and also R itself, in order to minimize c But

$$c = \sum_{r=1}^{R} \omega^{-\varepsilon} \left\{ \frac{\alpha}{p\delta} \left[e^{-(\mu+\delta)t_r} - e^{-\mu t_{r-1} - \delta t_r} \right] + e^{-\delta t_r} \right\}, \tag{3}$$

$$= \sum_{r=1}^{R} \omega^{-\varepsilon} \{ \omega^{-1} [(\omega x_r)^{(\mu+\delta)/\mu} - (\omega x_{r-1})(\omega x_r)^{\delta/\mu}] + (\omega x_r)^{\delta/\mu} \}, \tag{4}$$

$$= \sum_{r=1}^{R} x_r^{\varepsilon} (x_r - x_{r-1} + 1) \tag{5}$$

---

[1] If the optimum output or capacity needs to be reduced at some time period, the disinvestment decision may involve "premature" scrapping, which would add unnecessary complication to the analysis For a firm or industry facing a stable demand

The boundary conditions can be established by considering the derivative

$$\frac{dc}{dx_R} = (1 + \varepsilon)x_R^{\varepsilon} + \varepsilon x_R^{\varepsilon - 1}(1 - x_{R-1}). \tag{6}$$

Now, if $x_{R-1} \leq 1$, then $(dc)/(dx_R) > 0$ for all $x_R > 0$, so the optimum choice of x_R must be zero, and hence $t_R = \infty$ in effect. But $t_{R+1} \rightarrow \infty$, so t_R is too large. On the other hand, if $x_{R-1} > 1$, the optimum value of x_R is positive. Notice that, if $x_R > 1$ and if we include an extra replacement and an extra variable x_{R+1}, then the optimum value of x_{R+1} would also be finite. So, if $x_R > 1$, there are too few replacements Therefore, we have the following boundary conditions $x_{R-1} > 1$, $x_R \leq 1$, with x_0 fixed by history and $x_{R+1} = 0$ [2]

The conditions for a minimum present value of costs are given by $(\partial c)/(\partial x_r) = 0$, which implies that

$$x_{r-1} = 1 + (1 + \varepsilon^{-1})x_r - \varepsilon^{-1}x_r\left(\frac{x_{r+1}}{x_r}\right)^{\varepsilon} \tag{7}$$

for $r = 1$ to $R - 1$ and

$$x_{R-1} = 1 + (1 + \varepsilon^{-1})x_R, \tag{8}$$

since $x_{R+1} = 0$ Thus the optimum sequence of replacements can be found by solving the difference equation

$$x_{r-1} = 1 + (1 + \varepsilon^{-1})x_r - \varepsilon^{-1}x_r\left(\frac{x_{r+1}}{x_r}\right)^{\varepsilon}, \quad (r = 1 \text{ to } R), \tag{9}$$

subject to the boundary conditions x_0 given, $x_{R+1} = 0$, $0 < x_R \leq 1$

This equation can be used to establish that successive machine lives, as they are replaced, are continually rising Let

$$x_{r-1}x_r^{-1} = y_{r-1} = e^{\mu(t_r - t_{r-1})} = e^{\mu u_{r-1}}, \tag{10}$$

where $u_{r-1} = t_r - t_{r-1}$ is the life of the machine installed at time t_{r-1} Then equation (9) implies that

$$y_{r-1} = x_r^{-1} + 1 + \varepsilon^{-1} - \varepsilon^{-1}y_r^{-\varepsilon} \tag{11}$$

---

curve and costs which either decline or remain constant over time, the assumption of no decline in capacity would seem reasonable An assumption of constant capacity is made by Williamson (1971)

[2] The boundary condition implies that the machine installed at time t_R is never replaced, which contradicts assumption *b* of a constant machine life made by Williamson (1971, 1974) Smith (1966, pp. 141–49) has also developed a model of embodied technical change, in which he also assumes a constant machine life but recognizes that his approach is only an approximation

If y_{r-1} is subtracted from y_r, the result is

$$y_r - y_{r-1} = x_{r+1}^{-1} - x_r^{-1} - \varepsilon^{-1}(y_{r+1}^{-\varepsilon} - y_r^{-\varepsilon}) \tag{12}$$

But $x_{r+1} < x_r$, so $x_{r+1}^{-1} - x_r^{-1} > 0$ Also,

$$y_R - y_{R-1} = x_{R+1}^{-1} - x_R^{-1} > 0, \tag{13}$$

using equation (12), since $y_R^{-\varepsilon} = 0$, $\varepsilon > 0$, so that $y_R^{-\varepsilon} - y_{R-1}^{-\varepsilon} < 0$. Therefore, by induction, $y_r - y_{r-1} > 0$, and thus $e^{\mu u_r} - e^{\mu u_{r-1}} > 0$ implies $u_r > u_{r-1}$, which was to be established This result, that economic lives are rising as a result of the embodied technical change, is obvious in the case of the last replacement, since $y_R \to \infty$, and therefore $u_R \to \infty$, which means that the last machine in the chain is not replaced

The intuitive explanation for this result, that successive machine lives are rising, is simple Physically machines last forever and do not require maintenance, so that the only reason for replacing a machine is the saving in labor costs that is entailed in using the most technically advanced machine But, since technical advance saves only labor, the savings to be obtained with replacement become progressively smaller in relation to the fixed capital charge associated with replacement As labor costs become negligible in time, these savings are less than the capital charge, so that eventually no replacement is called for

The comparative dynamics of the model are such that machine lives will rise if the price of a machine increases or the wage rate falls To establish this result, which confirms Habakkuk's (1962, p 59) intuition, consider the sign of the total derivative of equation (11) with respect to ω.

$$\frac{dy_{r-1}}{d\omega} = x_r^{-1}\omega^{-1} - \varepsilon^{-1}\frac{dy_r^{-\varepsilon}}{d\omega} \tag{14}$$

But $y_{R-1} = x_R^{-1} + 1 + \varepsilon^{-1}$, since $y_R^{-\varepsilon} = 0$, and therefore $y_{R-1}^{-\varepsilon} = (x_R^{-1} + 1 + \varepsilon^{-1})^{-\varepsilon}$, with

$$\frac{dy_{R-1}^{-\varepsilon}}{d\omega} = -\varepsilon(x_R^{-1} + 1 + \varepsilon^{-1})^{-(\varepsilon+1)}x_R^{-1}\omega^{-1} < 0 \tag{15}$$

Hence $(dy_{R-1})/(d\omega) > 0$, and by induction $(dy_{r-1})/(d\omega) > 0$ Therefore $(du_{r-1})/(d\omega) > 0$, since $(du_{r-1})/(dy_{r-1}) > 0$ from equation (10) Consequently, the result is proved, since $\omega = (p\delta)/[b(0)u]$ with $(d\omega)/(dp) > 0$ and $(d\omega)/(dw) < 0$.

In general, numerical techniques are required to solve the difference equation (7); however, some insight into the general solution can be obtained by a consideration of the analytical solution to the special case when $\varepsilon = \mu/\delta = 1$. The general solution to the difference equation

$$x_{r-1} = 1 + 2x_r - x_{r+1} \tag{16}$$

is the quadratic form

$$x_r = a + br + \tfrac{1}{2}r^2. \tag{17}$$

Given the boundary conditions x_0 and $x_{R+1} = 0$, equation (17) becomes

$$x_r = \left(\frac{x_0}{R+1} - \frac{1}{2}r\right)(R + 1 - r). \tag{18}$$

Since $0 < x_R \leq 1$, we have $0 < x_0/(R+1) - \tfrac{1}{2}R \leq 1$, so that $\tfrac{1}{2}R(R+1) < x_0 \leq \tfrac{1}{2}(R+1)(R+2)$ or

$$g(R) < x_0 \leq g(R+1), \tag{19}$$

where $g(R) \equiv \tfrac{1}{2}R(R+1)$. For a given initial condition x_0, the inequality (19) determines the integer R uniquely Using equation (18), the ratio $y_r = (x_r/x_{r+1})$ can be evaluated, and it is easily confirmed that $(dy_r)/(d\omega) > 0$ and $(du_r)/(dw) < 0$, as must be the case since $(du_r)/(d\omega) < 0$ holds in the general case $\varepsilon > 0$.

II. Quantitative Analysis of the Early New England and *British Textile Industry*

In this section, the model developed previously is applied to the nineteenth-century New England and British textile industries The data for the analysis of the New England cotton textile industry is provided by Sinnott (1968, p 136) The labor-capital ratio $b(0)$ in 1830 was 0 045 operators per spindle, and by 1850 the figure had fallen to 0 026, which represents an annual average rate of Harrod-neutral technical change $\mu = 0.028$ At the same time, there was a fall in the number of spindle hours, and hence labor hours,[3] required per pound of cloth from 44 in 1830 to 32 in 1850. This change represents a Hicks-neutral annual avearge rate of technical advance of 0.016 in addition to the labor-saving advance, about which the present study is concerned [4] This decline in spindle hours required per pound will be neglected in the computation of the optimum scrapping dates and machine lives According to Sinnott (1968, pp. 140, 216), the interest rate $\delta = 0$ 06 remained constant over the period, while the (average) price p of a new machine fell from \$28 to \$18,[5] and

---

[3] This follows because the change in the number of operators per spindle, i e, the change in $b(t)$ between 1830 and 1850, has already been taken account of

[4] The neglect of the Hicks-neutral technical-change component of the overall technical advance is justified on the grounds that both the original Habakkuk argument and Williamson's model are concerned entirely with pure labor-saving technical advance, so that to introduce an entirely different form of technical change would be a distraction from the major issue

[5] McGouldrick (1968, p 240) has estimated machinery costs per spindle for a different sample of machines. His series shows a decline from \$27.97 in 1827 to \$19 93 in 1830 and \$11.74 in 1850 Thus McGouldrick's sample shows a fall which is greater in proportionate terms but slightly less in absolute terms between 1830 and 1850 than Sinnott's. Most of McGouldrick's other data is not comparable with Sinnott's and in any case is not suitable for use in the model. Thus it was thought better to take all the data from one source so as at least to provide a greater chance of internal consistency

TABLE 1

SEQUENCE OF ECONOMIC LIVES AND AGE AT REPLACEMENT

SINNOTT				WILLIAMSON				SINNOTT	
U S.		U K		U S		U K		U S	
1*	2†	3*	4†	5*	6†	7*	8†	9*	10†
1830		1830		1830		1830		1850	
1850	20	1853	23	1858	28	1887	47	1889	39
1880	30	1891	38	∞	∞	∞	∞	1942	53
1936	136	∞	∞					2032	90
∞	∞							∞	∞

* Date of installation of machine
† Age at replacement in years

TABLE 2

ACTUAL LIVES (U S)

Dates	Years
1830-40	15-20
1841-80	20-30
1881-90	25-30
1891-1915	30-35

SOURCE —Sinnott 1968, p 169

the annual wage rate rose from $180 to $197 over the 20-year period. Since the model (following Williamson) assumes static machine price and wage expectations, it is the initial 1830 values which are relevant for the present study. These figures imply a value of $1/\omega = 4.83$ for the New England industry in 1830, falling to 4.744 in 1850.

Numerical methods were used to obtain an (approximate) solution to the difference equation (7) above, based on Sinnott's 1830 data, and the corresponding introduction and replacement dates with machine ages at replacement are shown in the first two columns of table 1. These theoretical economic lives may be compared with the actual machine lives estimated by Sinnott (1968, p. 169) presented in table 2.

The theoretical replacement ages using 1830 data are remarkably close to the actual replacement ages for the sequence of the first two machines installed and scrapped. Of course the implied lives for machines installed in 1880 or subsequently have no basis in reality, but then the assumption of no physical deterioration must become critical once economic lives exceed 30 years. Coincidentally, the model is actually able to predict correctly the historical lengthening of machine lives without the need to reduce the expected rate of Harrod-neutral technical change. This contrasts with the models of Smith (1966) and Williamson (1971), which *assume* constant machine lives.

The model was applied to the United Kingdom textile industry in order to compare Anglo-American experience in replacement policy. Perhaps the most obvious difference between the two countries was the considerably lower wage rates paid in the United Kingdom. Rosenberg (1967), for example, quotes rates of \$1.02 per day for common mule spinners in England and \$1 08–\$1.40 in the United States during the 1820s In this study, Sinnott's 1830 data is retained apart from the assumption that the wage rate w was 20 percent lower in the United Kingdom, ensuring a lower value of $1/\omega = 3.857$ [6] It would seem reasonable that relatively free trade in investment goods such as textile machines would ensure fairly similar machine prices and rates of technical advance embodied in new machines between the two countries [7] Also capital movements should have ensured relatively similar interest rates

The results of the quantitative analysis are shown in columns 3 and 4 of table 1 Machine lives have increased simply as a result of the lower wages of unskilled workers, which is consistent with Habakkuk's (1962, p 56) qualitative evidence.

Williamson (1971, 1974) carried out a study of the United States and United Kingdom industries for the 1830s For the United States, he assumed a value[8] of $p/\alpha = 7\ 44$, which is exactly twice the figure of 3 62 for 1839, based on data and sources which Williamson himself (1971) quotes.[9] Sinnott's estimate of 3 46 for 1830 (or 3 8 for 1839) compares closely with this alternative estimate In fact, the data assumed by Williamson seem to be a little unreasonable when compared with the historical evidence [10] He also assumed a rate of Harrod-neutral technical advance $\mu = 0.05$,[11] which is also about twice Sinnott's estimate, and an interest rate $\delta = 0\ 08$ For the United Kingdom, he assumed a lower value of μ, 0 02, and an interest rate $\delta = 0\ 05$, while remaining data was the same as for the United States Columns 5–8 of table 1 show the replacement sequences and economic lives based on Williamson's assumptions. Lives of 28 and 47 years, respectively, for the United States and United Kingdom for the first machine in the chain compare with Williamson's (1971, p 1329, 1974, p 639) results of *constant* lives of 17 and 23, respectively Note that, even setting aside the questionable nature of Williamson's data and adopting his own data, there is still a considerable discrepancy between the results obtained here and his earlier 1971 results This is due to the contradictory nature of the assumptions embedded in his model Consequently, no support is found for his (1974) contention that his conclusions require little revision once his errors are corrected

---

[6] It is possible that the annual wage differential exceeded 20% if U K labor worked fewer hours than its U S. counterpart, which in the 1830s worked 12 hours a day, 6 days a week, according to Sinnott (1968, pp 130, 136) Nonetheless, the appropriate comparison should be between the hourly or daily rates, since these best reflect the difference between the cost of labor services in the two countries

[7] I would not wish to defend too strongly the hypothesis that there was equalization of machine prices between the United Kingdom and the United States prior to the lifting of a partial ban on exports of machines from the United Kingdom to the United States in 1843 However, the case for equalization is argued strongly by Williamson (1971, p 1329), and it is the assumption he makes throughout his analysis

[8] This figure is obtained indirectly for $p/\alpha = (K/X)^*/\alpha_L = [(pK)/(PX)]/[(wL)/(PX)]$, where $(K/X)^*$ = the capital-output ratio in value terms, α_L = labor's share, X = output of cloth in yards, and P = price of cloth per yard (in dollars) $(K/X)^*/\alpha_L$ is assumed to be equal to 3 0 (Williamson 1971, p 1328), and α_L assumed to be equal to 0 403 (Williamson 1974, p 639). No evidence is presented in support of these assumptions

[9] For 1839, $p = \$13\ 61$, $P = \$0\ 1175$ (McGouldrick 1968, pp 240, 145), and $X/K = 224$ (actually for Massachusetts in 1837 and quoted by Williamson [1971, p 1323]), so that $(K/X)^* = 0.5171$. Also $wL = \$0.0168$ per yard (McGouldrick 1968, p 147), and, using P, $\alpha_L = wL/PX = 0\ 143$, so that $p/\alpha = wL/pK = 3.62$

[10] Williamson's results (1971) are based on $(K/X)^* = 3\ 0$, which is over 600% higher than the corresponding figure based on his sources of 0 5171 Even if $(K/X)^*$ was supposed to be in terms of value added, the actual figure rises to only 0.75 (value added per yard = \$0 0808 [McGouldrick 1968, p 145]). Moreover, the historically observed value of α_L is 0.143 from Williamson's souces,r while Williamson's figure is nearly 300% greater. In value-added terms, the discrepancy is smaller, since α_L rises to 0.208

The final empirical study shows the changes which took place between 1830 and 1850 in the New England textile industry Sinnott's (1968, p 136) data indicate that by 1850 $1/\omega$ had fallen to 4 744 The most drastic change was a fall in the rate of Harrod-neutral change to less than 1 percent over the next 20 years However, a value of $\mu = 0\ 01$ is assumed for the purposes of the analysis The results are shown in columns 9 and 10 of table 1

This sequence of machine lives brings out some weaknesses of the model in a period of slow technical advance Obsolescence is the only explanation for scrapping in the model, and hence machine lives become considerable in such circumstances A more realistic but complex model would take account of repair costs rising with age and physical deterioration Initial machine costs and hence price might also vary with the technical durability built into the machine

Williamson (1971, pp 1328–32) devotes a great deal of attention to the proposition that U S tariffs on consumption goods which raised the price of textiles in the United States relative to the United Kingdom would shorten the economic lives of U S equipment Insofar as relative factor prices between the two countries are unaltered by the tariff,[12] this proposition is false Replacement dates and economic lives are purely the result of cost-minimizing behavior. In the absence of any scale effects,[13] product price has no bearing on replacement policy, as long as the capacity is required. This is not to imply that product price which influences the scale of the industry has no effect on decisions to *install* new equipment There is no logical connection between a decision to install a new machine and to scrap or replace an old one A failure to scrap at the time a new machine is introduced implies no more than an increase in productive capacity.

---

[11] Coincidentally, Williamson's assumed value for μ is not all that much greater than the sum of Sinnott's labor-saving advance of 2 8% plus the Hicks-neutral advance of 1 6%, so that the main difference between Sinnott and Williamson would seem to be in Williamson's neglect of the Hicks-neutral component, which should have been subtracted from the 5% figure

[12] Williamson (1974, p 639) does in fact raise the possibility that tariffs might have had some effect on wages in the United States However, this possibility seems to be discounted by the general equilibrium model of Pope (1972, p 413), who found that the return to labor never altered by more than 0 5% in any of the simulations of tariff changes. In Williamson (1971), the effect of the tariff on machine lives is due entirely to the rise in price of textiles, with no mention of any effect on wages

A related objection can be raised to Temin's (1966) thesis that the decline in the British steel industry from the 1880s to World War I was due to a slow growth in the domestic market At the same time, the restrictive tariff policies in the United States and Germany enabled these countries to obtain a younger and technologically superior capital stock by ensuring a rapid growth in the domestic market of these countries Temin concludes that, as a consequence, these countries were eventually able to gain a cost advantage over Great Britain in world markets[14]

While it is undoubtedly true that a rapid growth in productive capacity implies that the average machine age will be lower and, consequently, the capital stock will be biased toward more modern technology, it does *not* follow as a point of logic that such a bias gives any advantage in world markets Temin's thesis not only confuses average costs over the entire capital stock with marginal costs, which are the only relevant costs, but also has the peculiar implication that a firm is worse off because it is in an old, slow-growing industry and inherits a plant from the past, while a firm in a newer industry is better off because it has no (or few) plants from the past to inherit!

The relevant cost to a producer participating in any market is not some accounting measure of average costs of production taken over his entire capital stock but, rather, his marginal production costs represented by unit production costs with the latest and most superior technology The capital value of machines which are still economic and hence retained in use but embody obsolete technology will be written down i value to accord with the changed circumstances Only operating costs o an installed machine with negligible scrap value are relevant for it continued operation. It is this factor which enables such machines t compete on more equal terms with machines embodying the lates technology which have a rental or capital cost component as well as a operating cost component

---

[13] By "scale effects" are meant economies or diseconomies of scale which are exclud by assumption

[14] McCloskey (1973, pp. 11, 105–13) has successfully argued that the empiric magnitude of this so-called cost advantage was insignificant in the case of the British ste industry. However, in view of my subsequent argument, it would seem that the rebutt is superfluous in any case, as long as British entrepreneurs did not fall into the trap basing *marginal* decisions on *average* costs

III. Conclusion

In conclusion, this study has given at least some credence to Habakkuk's (1962) implicit model, which explains differences in observed replacement policy between the United States and the United Kingdom in terms of a difference in wage rates At the same time, it has shown how inconsistencies in Williamson's 1971 and 1974 models have given rise to a number of faulty predictions concerning Anglo-American differences in replacement policies A consistent solution to the problem posed by Williamson and other authors has been forthcoming Finally, a closely related thesis put forward by Temin has been shown to have no foundation

References

Bliss, C J. "On Putty-Clay " *Rev Econ Studies* 35 (April 1968) 105–32

Denslow, D , Jr , and Schulze, D "Optimal Replacement of Capital Goods in Early New England and British Textile Firms A Comment " *J P E* 82, no 3 (May/June 1974) 631–37

Habakkuk, H J *American and British Technology in the Nineteenth Century* New York Cambridge Univ Press, 1962

McCloskey, D N *Economic Maturity and Entrepreneurial Decline, British Iron and Steel, 1870–1913* Cambridge, Mass Harvard Univ Press, 1973

McGouldrick, P F *New England Textiles in the Nineteenth Century, Profits and Investment* Cambridge, Mass Harvard Univ Press, 1968

Pope, C "The Impact of the Ante-Bellum Tariff on Income Distribution " *Explorations Econ Hist* 9 (Summer 1972) 375–421

Rosenberg, N "Anglo-American Wage Differences in the 1820's " *J Econ Hist* 27 (June 1967) 221–29

Sinnott, T W "Investment Policies, Growth and Profitability in the New England Cotton Textile Industry, 1830–1914 " Ph D dissertation, Yale Univ , 1968

Smith, V L *Investment and Production* Cambridge, Mass Harvard Univ Press, 1966

Temin, P "The Relative Decline of the British Steel Industry, 1880–1913 " In *The Industrialization of Two Systems*, edited by H Rosovsky Cambridge, Mass Harvard Univ. Press, 1966

Williamson, J G "Optimal Replacement of Capital Goods The Early New England and British Textile Firm " *J P E* 79, no 6 (November/December 1971) · 1320–34

——— "Optimal Replacement of Capital Goods in Early New England and British Textile Firms Reply " *J P E* 82, no 3 (May/June 1974) 638–40

Some More Problems with Income-contingent Loans: The Case of Medical Education

Roger Feldman

University of North Carolina at Chapel Hill

In a recent article in this *Journal*, Nerlove (1975) notes that income-contingent loans for financing higher education may affect the borrower's income profile This is true both ex ante and ex post ex ante, because "adverse selection" loads the plan with poor insurance risks, ex post, because the "moral hazard" of repayments alters the borrower's choice between income and leisure Nerlove regrets that "unfortunately there is very little empirical evidence on these profiles and effects " In this comment I will present evidence on the moral hazard created by income-contingent loans for financing medical education [1]

Awareness that medicine is a secure and remunerative career, coupled with rising costs and tight public budgets, has fostered proposals for placing more of the burden on the student.[2] It is argued that the obligation should be financed by income-contingent loans *because* of moral hazard By taking income[3] the plan changes the earnings profiles of alternative medical specialties to favor the low-income primary care fields Challoner (1974) considers this a positive feature of the plan "As a result of the pooled risk or insurance principle, no borrower would have to worry about a large debt that he could not repay Thus, his decision for a higher or lower paying professional activity need not be colored by the specter of a large conventional debt, soon to be repaid This scheme could have important effects on specialty distribution " Since federal policy encourages training more primary care physicians (Kennedy 1974), income-

---

[1] Adverse selection can be ignored if the student pays the *full cost* of education Full cost, or average annual cost per student, was estimated by Hanft et al (1974) at $12,650 in 1972–73. Few students could finance this large obligation for 4 years without the loan, so adverse selection is eliminated, and I concentrate on the problem of moral hazard

[2] The state of New Jersey is considering a tuition increase at the College of Medicine and Dentistry from $1,750 to $5,000 per year

[3] "Taxing income" and "tax rate" only suggest the analogy to the federal income tax The loan repayment rate would not be added onto the federal rate. But neither should repayments be deductible from taxable income A deduction is a subsidy and defeats the full cost loan concept.

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved

contingent loans may have beneficial side effects by favoring these low-income specialties.

However, income-contingent loans change the income-leisure trade-off for all physicians. A plan which uses a flat repayment rate will reduce weeks worked if individual supply curves of labor slope upward. A plan which creates a progressive tax by exempting some income from taxation must levy a higher rate on taxable income to raise as much revenue as the flat rate version.[4] Weeks worked will fall more under this second plan

Reducing physicians' weeks of work is undesirable because a given number of physicians produces fewer services. Therefore, we have a classic economic problem: income-contingent loans may encourage mor primary care physicians, but only at the cost of reducing total physicians services Can the size of this trade-off be estimated before an actual pla is adopted? I argue that it can and offer the following evidence.

From a national survey[5] of individual physicians (American Medic Association 1970) I calculated the average yearly net practice incom for eight specialties[6] Few physicians were surveyed at some ages in som specialties, so there was considerable sampling variation. I smoothed t income series by fitting fourth-degree polynomials in age with ordina least squares regressions. Pretax present value of income for sev specialties relative to general practice was computed from the adjust data.

Next, consider the full cost of medical school The student loan $12,650 interest free per year would owe $50,600 after 4 years (assum no repayments during school) But an interest-free loan subsidizes student, because the plan itself must borrow funds Assume that the p borrows at 10 percent interest and passes these costs on to the stud who owes principal and interest of $64,579 after 4 years[7] For the p to break even it must collect revenue with present value of $64,579 fr the average graduate.

First, I examine a plan which levies a flat repayment rate on all fu practice income Using the income data and the proportion of physicians going into each specialty (American Medical Associa 1971), I created a hypothetical average physician. Different rates ap against his income created profiles of repayments. Discounting t

---

[4] Strauss (1970) proposed this plan

[5] Questionnaires (with six follow-ups) were mailed to a national probability sam physicians stratified by practice type Questions pertained to the year 1970 The response rate of 57 percent produced a data base of 5,084 physician records, bu responses eliminated 2,174

[6] General practice (GP), internal medicine, pediatrics, other medical sp general surgery, obstetrics and gynecology (OBG), other surgical specialty, other sp

[7] Although 10 percent is a reasonable interest rate for 1976, future credit conditions will change the present value of the loan. The present value can be h stant by using variable interest rates to compute repayments.

TABLE 1

EFFECT OF 20% INCOME-CONTINGENT TAX ON PRESENT VALUE OF INCOME AT 6% AND 10% DISCOUNT RATES

SPECIALTY	PRESENT VALUE OF INCOME ($) RELATIVE TO GENERAL PRACTITIONER				CHANGE IN PRESENT VALUE OF INCOME	
	Current Financing		Income-contingent Loan Plan			
	6%	10%	6%	10%	6%	10%
Internal med	−48,361	−63,730	−38,703	−50,988	9,658	12,742
Pediatrics	−95,494	−91,014	−76,393	−72,810	19,101	18,204
Other med	−63,148	−73,401	−95,809	−84,638	−32,661	−11,237
Gen surgery	26,510	−29,015	21,209	−23,211	−5,301	5,804
Other surgery	78,081	9,066	62,466	7,254	−15,615	−1,812
OBG	32,669	−12,969	26,135	−10,375	−6,534	2,594
Other	−5,813	−39,724	−4,649	−31,779	1,164	7,945

SOURCES —Physicians' incomes from American Medical Association (1970), full cost of medical education, used to compute present values under income-contingent loan plan, from Hanft et al (1974)

profiles back to the end of medical school at 10 percent, I found that the plan could break even at repayment rate of 20 percent

I applied the break-even rate to income profiles of each specialty and recomputed present values of income, and the changes in present values, caused by the plan The changes in present values were multiplied by Sloan's (1970) estimate of specialty choice[8] to find the change in residents choosing each specialty Dividing this number by the average length of residency[9] produced the change in the net flow of physicians per year

Results are shown in tables 1 and 2 In table 1 discount rates of 6 and 10 percent illustrate how the results change under different assumptions, in table 2 only the 10 percent rate is used to calculate the effect on specialty choice.

In table 1 the present value of income for six specialties is negative at the discount rate of 10 percent Specialists do not earn extraordinary incomes relative to GPs

At a 6 percent discount rate the three surgical specialties (general, other and OBG) become less attractive relative to general practice, the two low-income primary care specialties (internal medicine and pediatrics) are more attractive At a 10 percent discount rate the change in present value becomes positive for general surgery and OBG Therefore, predictions in table 2 depend crucially on the choice of a discount rate This is a

---

[8] Sloan found a small but significantly positive relation between the number of residents choosing each specialty and the present values of income

[9] I assume a 1-year internship in general practice followed by a 3-year residency in all specialties. This was approximately correct in 1970, but recent changes (especially the development of 2-year residencies in family practice) suggest caution in interpreting my estimates.

TABLE 2

EFFECT OF 20% INCOME-CONTINGENT TAX ON SPECIALTY CHOICE

Specialty	Change in Residents	Change in New Physicians per Year	New Physicians per Year Current Financing	% Change New Physicians per Year
Internal med.	96	32	2,074	1.5
Pediatrics	137	46	640	7.2
Other med .	−84	−28	267	−10.5
Gen surgery	44	15	1,472	1.0
Other surgery	−14	−5	1,962	−0.03
OBG	19	6	517	1.2
Other .. .	60	20	3,220	0.6
Gen. practice		−86	828	−10.4

SOURCES.—New physicians per year, current financing, from American Medical Association (1971), specialty choice coefficients from Sloan (1970)

common problem with empirical studies of behavior over time. One solution is to use the market rate of interest, and 10 percent is a reasonable estimate in a period of high interest rates.

The first dimension of moral hazard is the effect on specialty choice. Other medical specialties show the largest change, minus 10 percent, but general surgery, other surgery, OBG, and other specialties all change by less than 2 percent. The sum of the changes in seven specialties must equal the change in general practitioners, because the plan does no increase the number of medical school graduates. Eighty-six fewe physicians will choose general practice. The number of GPs and primar care specialists falls by eight. Therefore, the plan will not influenc specialty choice toward policy goals. Changes will be small, and the may be unexpected.

The second dimension of moral hazard is the change in physician work effort. The effect of a 20 percent income-contingent tax on wo effort depends on the supply response of physicians to lower wage Studies of physicians' work effort show a small negative response wh wages and work are measured by the week (Sloan 1973, 1974, 197 Supply elasticities range from about 0.03 to 0.25. For a representati estimate of 0.10, the supply elasticity is (dweeks/weeks)/(dwage/wage) 0.10. Under this plan dwage/wage = −0.20, so dweeks/weeks = −0 Of the estimated 10,980 new physicians in 1970 who were graduates U.S. and Canadian medical schools (American Medical Associat 1971), the loss is 220 physicians—the classes of two average-size school

---

[10] Several qualifications are necessary. If the plan greatly alters specialty choice, break-even repayment rate will change. The rate rises if fewer physicians choose h income specialties. Income in these now-scarce specialties would probably rise, offse the initial specialty change and damping the increase in the repayment rate. If all p cians work less, the rate also rises. In this study first round changes were small enou make an iterative solution unnecessary

TABLE 3

EFFECT OF EXEMPTION AND 34% INCOME CONTINGENT TAX ON PRESENT VALUE OF INCOME AT 6% AND 10% DISCOUNT RATES

SPECIALTY	PRESENT VALUE OF INCOME ($) RELATIVE TO GENERAL PRACTITIONER				CHANGE IN PRESENT VALUE OF INCOME	
	Current Financing		Income-contingent Loan Plan			
	6%	10%	6%	10%	6%	10%
Internal med.	−48,361	−63,730	−29,074	−42,252	19,287	21,478
Pediatrics	−95,494	−91,014	−60,170	−60,257	35,324	30,757
Other med.	−63,148	−73,401	−38,824	−48,633	24,324	24,768
Gen. surgery	26,510	−29,015	20,353	−19,337	−6,157	9,678
Other surgery	78,081	9,066	54,389	5,797	−23,692	−3,269
OBG	32,669	−12,969	24,417	−8,747	−8,252	4,222
Other	−5,813	−39,724	−982	−26,405	4,831	13,319

SOURCE —Exempt income, from U S Bureau of the Census (1973, Table 249)

TABLE 4

EFFECT OF EXEMPTION AND 34% INCOME-CONTINGENT TAX ON SPECIALTY CHOICE

Specialty	Change in Residents	Change in New Physicians per Year	New Physicians per Year, Current Financing	% Change New Physicians per Year
Internal med	161	54	2,074	2 6
Pediatrics	231	77	640	12 0
Other med	185	62	267	23 2
Gen surgery	73	24	1,472	1 6
Other surgery	−25	−8	1,962	−0 4
OBG	32	11	517	2 1
Other	100	33	3,220	1 0
Gen practice		−253	828	−30 6

The second plan exempts part of physicians' income from taxation. I created an experimental income-contingent tax base by subtracting earnings of male college graduates (U S Bureau of the Census 1973, table 249) from physicians' income by age On the new base I calculated a break-even repayment rate of 0.34 Since some income is exempt, the rest must be taxed heavily

Tables 3 and 4 show how this plan changes specialty choice Table 3 resembles table 1. At a 6 percent discount rate primary care specialties become more attractive and surgical specialties less attractive relative to general practice. At 10 percent discount the change in present value of income reverses sign for general surgery and OBG For all medical specialists the progressive tax plan makes a large positive difference in present value of income.

This plan effectively influences physicians to choose pediatrics, other medicine, and, to a lesser extent, internal medicine. However, the increases come mainly from general practitioners, since surgical and other specialists are not greatly affected. General practitioners lose 253 and all primary care physicians are down 122. This plan, like the first one, mainly causes substitution within the primary care group

The exemption may cause this unexpected result. During the first few years of practice GPs earn large incomes while residents earn about as much as college graduates. Therefore, GPs are taxed while residents escape taxes. Since the discounting process strikes heavily at income earned in the near future, GPs may lose even though the specialists will be taxed in the far future.

Last, we find the effect of this plan on physicians' work effort. If the supply elasticity is 0.10, a 0.34 repayment rate will reduce work effort by 0.034. But the exemption also creates a pure income effect which can be evaluated by the following analysis. Net income equals gross income minus taxes, or $Y - r(Y - E)$, where Y is gross income, r is the 0.34 repayment rate, and E is the exempted income of an average male college graduate ($13,509 in 1970) Net income is $4,593 when Y is zero Estimates of the income elasticity of supply, (dweeks/weeks)/(dincome/income), vary from about −0.03 to −0.6 Choose a point estimate of −0.3. The average physician on the AMA survey earned $43,412 in 1970 Therefore, the income effect would reduce work effort by dweeks/weeks = −0.3($4,593/$43,412) = −0.032

The disincentives caused by the repayment rate and the exemption are approximately equal. Altogether work falls by 6.6 percent The loss is 725 new physicians per year, three times the loss from the first plan

Both plans transfer the cost of medical education to the student My findings show that neither plan encourages more physicians to choose primary care. Income-contingent loans, therefore, cannot satisfy two goals. Both plans also penalize work effort by taxing income The drawback is about three times as serious under the progressive tax plan This evidence favors a flat repayment rate.

References

American Medical Association. *Seventh Periodic Survey of Physicians*. Chicago. AMA Center for Health Services Res. and Development, 1970.

———. "Medical Education in the United States, 1970–71." *J. American Medica Assoc.* 218 (November 22, 1971): 1199–1286.

Challoner, David R. "The Medical Student Pays for His Education." *Ne England J. Medicine* 290 (January 17, 1974): 160–62.

Hanft, Ruth S., et al. *Costs of Education in the Health Professions*. DHEW Pub. no (HRA) 74–32. Washington: U.S. Dept. Health, Education, Welfare, 1974.

Kennedy, Edward M *Health Professions Educational Assistance Act of 1974* Senate Report 93-1133 (93d Congress, 2d session) Washington Government Printing Office, 1974

Nerlove, Marc "Some Problems in the Use of Income-contingent Loans for the Finance of Higher Education " *J P E* 83, no 1 (February 1975) 157–83

Sloan, Frank A. "Lifetime Earnings and Physicians' Choice of Specialty " *Indus and Labor Relations Rev* 24 (October 1970) 47–56

———. *Supply Responses of Young Physicians An Analysis of Physicians in Residency Programs.* Pub no. R-1131-OEO Santa Monica, Calif RAND, 1973

——— "A Microanalysis of Physicians' Hours of Work Decisions " In *The Economics of Health and Medical Care,* edited by Mark Perlman London Macmillan, 1974.

———. "Physician Supply Behavior in the Short Run " *Industrial and Labor Relations Review* 28 (July 1975) 549–69

Strauss, Maurice B "Financing Medical Education " *New England J Medicine* 283 (July 2, 1970). 45–46

U.S. Bureau of the Census *U S Census of Population 1970* Vol 1, pt 1, *Characteristics of the Population, U S Summary,* sec 2 Washington Government Printing Office, 1973.

Policy Implication of Full-Loss-Offset Capital Gains Tax on Forward Contracts: A Note on Leland's "Optimal Forward Exchange Positions"

Dan Galai

The Hebrew University, Jerusalem

In his paper "Optimal Forward Exchange Positions," Leland arrives at the following policy implication. "If the government wishes to encourage speculative trading, it should consider the imposition of a high tax with full loss offset" (1971, p. 262). His basic assumption is that no initial commitment of resources is needed in order to purchase a forward exchange contract. Other assumptions are that traders, who are risk averters, act to maximize their expected utility of terminal wealth and are price takers in spot or forward exchange markets.

In this note it is shown that, in the case of forward contracts for foreign currencies, where investors are allowed to buy and sell (even short) the underlying currencies, Leland's policy implication can be applied for a special case only. Moreover, the economic motivation for encouraging trading in speculative trading is absent.

I

The problem of a single exchange market with no alternative risky asset (model I in Leland's paper) may be stated as

$$\underset{x_1}{\text{maximize}}\ E\{U[(1 + r)B_0 + \tilde{e}_1 x_1]\}, \tag{1}$$

where U is the utility function, E is the expected value operator, B_0 is the initial wealth of the investor, $\tilde{e}_1$ is the profit (or loss) per unit of currency (or commodity) contracted (i.e., $\tilde{e}_1$ equals the spot unit price of the currency at maturity $\tilde{S}_1$ minus its exercise price or forward rate F),

---

The first draft of this note was written while I was a lecturer at the University of Chicago Graduate School of Business. I would like to thank F. Black, J. Frenkel, J. P. Gould, and H. Leland, who commented on an earlier version of the note. All remaining errors are mine.

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved

and x_1 is the number of units of the currency. (The title denotes a random variable)

If x_1^* is the solution to (1), then the solution to a similar problem, with a proportional change λ in $\tilde{e}_1$, will be $x_1^{**} = x_1^*/\lambda$ (Leland 1971, p. 261, eq. [16]) Therefore, when $0 < \lambda < 1$, $x_1^{**} > x_1^*$; that is, the number of units of the contracted commodity will increase if a proportional reduction of profits is expected This result, according to Leland, "...holds when there are multiple exchange markets and other risky assets. Furthermore, by selective taxation the government may affect certain markets while leaving others totally unaffected" (p 262)

Allowing for multiple forward exchange markets and other risky assets (Leland's model III, p 267, equation [34]), the problem is

$$\underset{\underline{x}_1,\underline{x}_2}{\text{maximize}}\, E\{U[(1 + r)B_0 + \tilde{\underline{e}}_1\underline{x}_1 + (\tilde{\underline{e}}_2 - r\underline{p}_2)'\underline{x}_2]\}, \qquad (2)$$

where $\underline{x}_1$ and $\underline{x}_2$ are vectors of investment in forward (or futures) contracts and in risky assets, respectively, $\underline{p}_2$ is the vector of prices of risky assets, and $\tilde{\underline{e}}_1$ and $\tilde{\underline{e}}_2$ are random vectors of net return per unit of investment in forward (futures) contracts and risky assets, respectively Imposing a capital gains tax, with full loss offset, of $(1 - \lambda_1^i)$ on profit $\tilde{e}_1^i$ from the ith contract, " the new optimal solution is the former one divided by λ_1^i"(1971, p 267)

Nevertheless, the *total* demand for "speculation" will not change in the case that full-loss-offset tax is imposed on the forward contract, and Leland's claim about encouragement of speculative trading by imposing the tax is void in that respect The mere fact that the number of contracts may have increased has no economic content per se

Under an assumption of zero transaction costs, the contract is a zero-sum game If a full-loss-offset tax of 50 percent on profits from forward contracts is imposed on buyers *and* sellers, then every contract can be converted to one that specifies twice as many units of currency as before This will leave everyone in the same position as before The government, which shares in profits and losses of both parties, will have zero net revenues from the tax. Because no initial outlay is required for trading forward contracts, there is no income effect from the tax [1] The buyers and sellers will maintain the same expected return and risk, and forward prices will remain unchanged

The analysis does not show how a country can profit from increasing the number of contracts to deliver foreign currency at maturity Actua delivery of the foreign currency is not needed, as positions can be close by the traders at maturity by offsetting transactions. If the seller of th forward currency wants to hold a (partially) covered position, then this should be incorporated into the model

---

[1] Musgrave (1959, chap 14) analyzed the case of a full-loss-offset tax on capita investments. He claimed the tax to have no substitution effect, only an income effe (p. 321). The investment he considered requires an initial outlay of resources; howeve he noted that "gambling gains may be taxed and *refunds disallowed*, if gambling is to b discouraged" (p. 320, emphasis added)

In the next section, I describe the equilibrium relationship between the forward and spot prices of the currency and incorporate both forward and spot positions into the model It will also be shown that Leland's result is a special case that will not hold if the tax on spot positions is different from the tax imposed on the interest rate

II

For forward contracts, it has been shown[2] that, in equilibrium, $F = So[(1 + r)/(1 + r_f)]$, where F is the forward rate (i e , the future price in domestic currency of a unit of foreign currency to be delivered at a predetermined date), So is the spot price of the foreign currency, and r and r_f are the interest rates in the domestic and foreign country, respectively If $F > So[(1 + r)/(1 + r_f)]$, then the investor can earn arbitrage profits He will borrow $So/(1 + r_f)$ dollars and will promise to pay back $[So/(1 + r_f)](1 + r)$ at time T With $So/(1 + r_f)$ dollars, he can buy $1/(1 + r_f)$ units of the foreign currency (say, pounds sterling) which can be invested to earn $[1/(1 + r_f)](1 + r_f) = £1$ at time T He will also sell a forward contract promising him F dollars for £1 he must deliver at time T At time T, the investor will have £1, which he will have to deliver to cover his forward position While he is fully covered in his foreign-currency obligations, he will have a dollar gain, as he receives F dollars while he owes only $So[(1 + r)/(1 + r_f)] < F$ This situation cannot continue, and the price of the forward contract will tend to go down or that of spot currency to go up Allowing for short sale of the foreign currency will cause the elimination of a possibility for $F < So[(1 + r)/(1 + r_f)]$ Thus an investor can either buy a forward contract or, equivalently, create on his own account by borrowing $So/(1 + r_f)$ (costing him $[So/(1 + r_f)]r$) and buying the foreign currency in the spot market. By the latter method, he has the same risk as the forward buyer, expects the same return, and in equilibrium it will cost him the same

III

To illustrate my argument, consider the simple case of one foreign currency (pounds sterling) and a forward contract for the same currency In this case, the investor's problem (after imposing the tax on profits from forward trading) is to

$$\underset{x_1, x_2}{\text{maximize}}\ E\left(U\left\{(1 + r)B_0 + \lambda_1 \tilde{e}_1 x_1 + \left[\tilde{e}_2 - r\frac{So}{(1 + r_f)}\right]x_2\right\}\right) \tag{3}$$

---

[2] See, e.g., Grubel (1966, p 5), Stoll (1966), or Frenkel and Levich (1975)

In equilibrium, $F = So[(1 + r)/(1 + r_f)]$; therefore

$$\tilde{e}_1 = \tilde{S}_1 - F = \tilde{S}_1 - So\frac{(1 + r)}{(1 + r_f)} = \tilde{S}_1 - \frac{So}{(1 + r_f)} - r\frac{So}{(1 + r_f)}, \quad (4)$$

where $\tilde{S}_1$ is the stochastic future spot price of the currency. But $\tilde{e}_2 = \tilde{S}_1 - [So/(1 + r_f)]$; thus $\tilde{e}_1 = \tilde{e}_2 - r[So/(1 + r_f)]$, and (3) becomes

$$\underset{x_1, x_2}{\text{maximize}}\, E\left(U\left\{(1 + r)B_0 + \left[\tilde{e}_2 - r\frac{So}{(1 + r_f)}\right](\lambda_1 x_1 + x_2)\right\}\right). \quad (5)$$

Expression (5) has the same solution as Leland's initial problem; therefore the optimal solution to (5)—$(\lambda_1 x_1 + x_2)^*$—will be equal to x_1^* of (1) Increasing x_1 by $1/\lambda_1$ is a valid and straightforward solution [3]

The solutions x_1^* to (1) or $(x_1 + x_2)^*$ to (5), when $\lambda = 1$, or $(\lambda_1 x_1 + x_2)^*$ to (5) $(0 < \lambda_1 < 1)$, are equal and therefore give the same optimal holding of a "risky-asset portfolio," either in its futures form or the equivalent, "self-made" forward contract Because of the way the problems were structured, one cannot decompose $(x_1 + x_2)^*$ or $(\lambda_1 x_1 + x_2)^*$ to their "optimal" components for an individual, and any combination of x_1 and x_2 which satisfies the optimal solution can be accepted [4]

Clearing of the market can be achieved, again, by increasing the number of contracts by $1/\lambda$ But note that the aggregate position in the spot market is unchanged Therefore total demand for "speculation" is unchanged, and no pressures on So or F are expected.

Problem (3) is a special case $(\lambda_2 = 1)$ of the following problem, for which Leland's policy implications do apply.[5]

$$\underset{x_1, x_2}{\text{maximize}}\, E\left(U\left\{(1 + \lambda_2 r)B_0 + \lambda_1 \tilde{e}_1 x_1 + \lambda_2\left[\tilde{e}_2 - \frac{rS_0}{(1 + r_f)}\right]x_2\right\}\right) \quad (6)$$

$$= E\left(U\left\{(1 + \lambda_2 r)B_0 + \left[\tilde{e}_2 - \frac{rS_0}{(1 + r_f)}\right](\lambda_1 x_1 + \lambda_2 x_2)\right\}\right). \quad (6')$$

---

[3] An individual, as a price taker, can adjust his position by increasing x_2 Hence increasing x_1 by $1/\lambda_1$ is not a unique solution for an individual. However, for the market as a whole, increasing $\sum_{i=1}^{n} x_{i1}$ by $1/\lambda_1$ will clear the market

[4] It is assumed that the tax is imposed on buyers and sellers, if it is imposed on buyers only, then the sellers quite probably will have a more risky position, while the government will hold the balance of the risky asset (i.e., its right to $[1 - \lambda]$ of buyers' realization)

[5] *Leland does not assume in any of his models the existence* of a tax on interest rate (payment or receipts). Here I show the class of cases to which his conclusions will apply, and it requires that the same tax (or no tax at all—as a private case) be imposed on interest rate as on the commodity. In order to be consistent, we must compare cases with the same tax (or no tax at all) on interest rate; the only variables we will control are the tax rate on profits (losses) from forward and spot currency.

Here, a tax (with full-loss-offset) of $(1 - \lambda_2)$ is imposed on the interest rate and (the same tax) on the return $\tilde{e}_2$ from the underlying commodity The original problem, (1), is a special case of (6), where $\lambda_1 = \lambda_2 = 1$ Comparing the solution to (6′) (i e , $[\lambda_1 x_1 + \lambda_2 x_2]^*$) with that of a similar problem but with $\lambda_1 = 1$ (with a solution $[x_1 + \lambda_2 x_2]^*$), we can conclude that the demand for x_1 and/or x_2 will increase, increasing x_1 by $1/\lambda_1$ is, again, a valid solution

IV

Now assume that a tax with full loss offset is imposed on all *risky* assets, so that $\lambda_1 = \lambda_2 = \lambda$ $(0 < \lambda < 1)$, then the problem becomes

$$\underset{x_1, x_2}{\text{maximize}}\, E\left(U\left\{(1 + r)B_0 + \lambda\tilde{e}_1 x_1 + \left[\lambda\tilde{e}_2 - \frac{rS_0}{(1 + r_f)}\right]x_2\right\}\right) \tag{7}$$

$$= E\left(U\left\{(1 + r)B_0 + \lambda\left[\tilde{e}_2 - \frac{rS_0}{(1 + r_f)}\right]x_1 + \left[\lambda\tilde{e}_2 - \frac{rS_0}{(1 + r_f)}\right]x_2\right\}\right) \tag{7′}$$

By buying an additional unit of contract, the investor has a right to $\lambda\tilde{e}_1 = \lambda\{\tilde{e}_2 - [(rS_0)/(1 + r_f)]\}$ at maturity, by creating the "self-made" contract, his claim "at maturity" is to $\lambda\tilde{e}_2 - [(rS_0)/(1 + r_f)]$ only. Note that no tax is imposed directly on the interest rate, but actually it is imposed on the implicit borrowing when buying futures, $\lambda\tilde{e}_1 = \lambda\{\tilde{e}_2 - [(rS_0)/(1 + r_f)]\} = \lambda\tilde{e}_2 - \lambda[(rS_0)/(1 + r_f)]$ Therefore buying a forward contract becomes relatively cheaper than creating the same position on one's own account[6] (where one has to pay the full interest rate), that is, $\lambda\tilde{e}_1 = \lambda\tilde{e}_2 - \lambda[(rS_0)/(1 + r_f)] > \lambda\tilde{e}_2 - [(rS_0)/(1 + r_f)]$

Rational investors will buy a contract and at the same time will go short on the currency and lend the proceeds from the sale Such a combination has a fixed, riskless yield per unit of the underlying commodity, which is not a function of the foreign currency's price at maturity Net return on a unit of a contract is $\lambda\tilde{e}_1 = \lambda\{\tilde{e}_2 - [(rS_0)/(1 + r_f)]\}$, the net return on a short position in currency is $-\lambda\tilde{e}_2$, and lending $S_0/(1 + r_f)$ will yield $(rS_0)/(1 + r_f)$ Thus the total return on one unit of portfolio is $\lambda\{\tilde{e}_2 - [(rS_0)/(1 + r_f)]\} - \lambda\tilde{e}_2 + [(rS_0)/(1 + r_f)] = [(rS_0)/(1 + r_f)] - \lambda[(rS_0)/(1 + r_f)] = [(rS_0)/(1 + r_f)](1 - \lambda)$ The term $[(rS_0)/(1 + r_f)] \times (1 - \lambda)$ is exactly the "tax subsidy" provided by the government to the forward buyer (at the expense of futures sellers—if the tax is imposed on them as well) by implicitly allowing him to borrow money cheaply

---

[6] One way to eliminate the possibility is to impose a tax (at a rate $1 - \lambda$) on the interest paid (or received) on loans that are needed in order to spot buy (or sell) the foreign currency. But then we are in the position described by eq (6) Leland's result will hold only if the tax on the interest is the same as the tax on $\tilde{e}_2$

It was shown that, by buying a right on one unit of foreign currency at a future period, selling short one unit in a spot market, and lending the proceeds of the short sale, one can increase his final wealth while taking *no additional* risk The optimal solution is, therefore, to invest as much as possible in x_1 while selling short x_2 and lending money Such a solution is not an equilibrium solution and will be eliminated immediately from the market by reducing expected profits on forward contracts or increasing them for spot commodities Either way, the reduction in the relative profits or forward positions will make them less attractive to investors than before, and we may find ourselves in Leland's case of model II of a shift of the return distribution by a constant, with the conclusion that demand for futures may very well fall!

A new equilibrium will be achieved at prices S_0' and F', so that $F' = [S_0'/(1 + r_f)][1 + (r/\lambda)]$ Then "self-made" forward positions will yield $\lambda\{\tilde{S}_1 - [S_0'/(1 + r_f)]\} - [(S_0'r)/(1 + r_f)]$, while ("original") contracts will have a net return of $\lambda(\tilde{S}_1 - F) = \lambda\{\tilde{S}_1 - [S_0'/(1 + r_f)] \times [1 + (r/\lambda)]\} = \lambda\{\tilde{S}_1 - [S_0'/(1 + r_f)]\} - [(S_0'r)/(1 + r_f)]$ For a given risk of $\tilde{S}_1$ and zero investment, the yields of the alternative investments are equal, and equilibrium is restored Again, the equilibrium is only partial and is with respect to the spot and forward prices General equilibrium will be restored only after the optimal quantities of x_1 and x_2 have been determined for *all* traders in the market This last point is especially important in the case in which the tax is imposed on one type of trader only, that is, buyers or sellers, or when differential taxes are imposed Then, increasing $\sum_i x_{1i}^*$ by $1/\lambda_B$ (where $1 - \lambda_B$ is the tax rate of buyers' profits from futures, and $\sum_i$ is summation over all buyers) would not necessarily be met by an equivalent increase in the supply, the latter will virtually tend to increase by $1/\lambda_S$ (where $1 - \lambda_S$ is the tax rate on sellers)

References

Frenkel, J A, and Levich, R M. "Covered Interest Arbitrage. Unexploited Profits?" *J.P E.* 83, no 2 (April 1975) 325–38.

Grubel, H G. *Forward Exchange, Speculation, and the International Flow of Capital* Stanford, Calif · Stanford Univ Press, 1966

Leland, H. P. "Optimal Forward Exchange Positions." *J.P.E* 79, no 2 (March/April 1971): 257–69.

Musgrave, R A *The Theory of Public Finance*. New York· McGraw-Hill, 1959

Stoll, H. "The Determinants of Forward Foreign Exchange Rate." Ph D dissertation, Univ. Chicago, 1966.

The Social Costs of Monopoly and Regulation: Some Comments

Göran Skogh

University of Lund, Sweden

Richard Posner (1975) has recently presented an interesting alternative to existing estimates of the social costs of monopoly and regulation His estimates are based on the notion that potential monopolists will be willing to expend up to the expected rents from a monopoly to obtain the monopoly so that under perfect competition these expected rents will be converted into social costs However, when it is profitable to establish a *cartel*, it will also be in the common interest of the members of the cartel to protect against violations of the cartel's rules as well as against threats from outsiders, and these protection costs are likely to affect the cartel's optimal price and quantity and thus its profits [1] In this note, I extend Posner's analysis to show that the existence of such protection costs can cause either a higher *or* lower price (than the simple monopoly price) to be optimal for the cartel

I. Posner's Model

Posner makes the following simplifying assumptions "(1) Obtaining a monopoly is itself a competitive activity, so that, at the margin, the cost of obtaining a monopoly is exactly equal to the expected profit of being a monopolist. (2) The long-run supply of all inputs used in obtaining monopolies is perfectly elastic (3) The costs incurred in obtaining a monopoly have no socially valuable by-products (1975, p 809)

Figure 1*a* is analogous to Posner's figure 2 and illustrates a cartelized market with constant marginal production costs [2] The price rise caused

---

I wish to thank the Bank of Sweden Tercentenary Foundation for financial support The ideas in this note have been presented to the law and economics workshop at University of Chicago. I am grateful to Gary S Becker, William M Landes, Richard A Posner, and Charles Stuart for discussions and helpful comments

[1] A "monopoly" might be defined as a single firm for which all internal and external marginal protection costs are zero I will use the word "cartel" when such internal costs are positive (for one or many firms)

[2] Posner assumes constant marginal production costs This is a strong assumption for an analysis of the short run On the other hand, in the long run when new producers can enter, the cartel may break down. I retain Posner's assumption for simplicity

[*Journal of Political Economy*, 1976, vol. 84, no. 6]

© 1976 by The University of Chicago. All rights reserved.

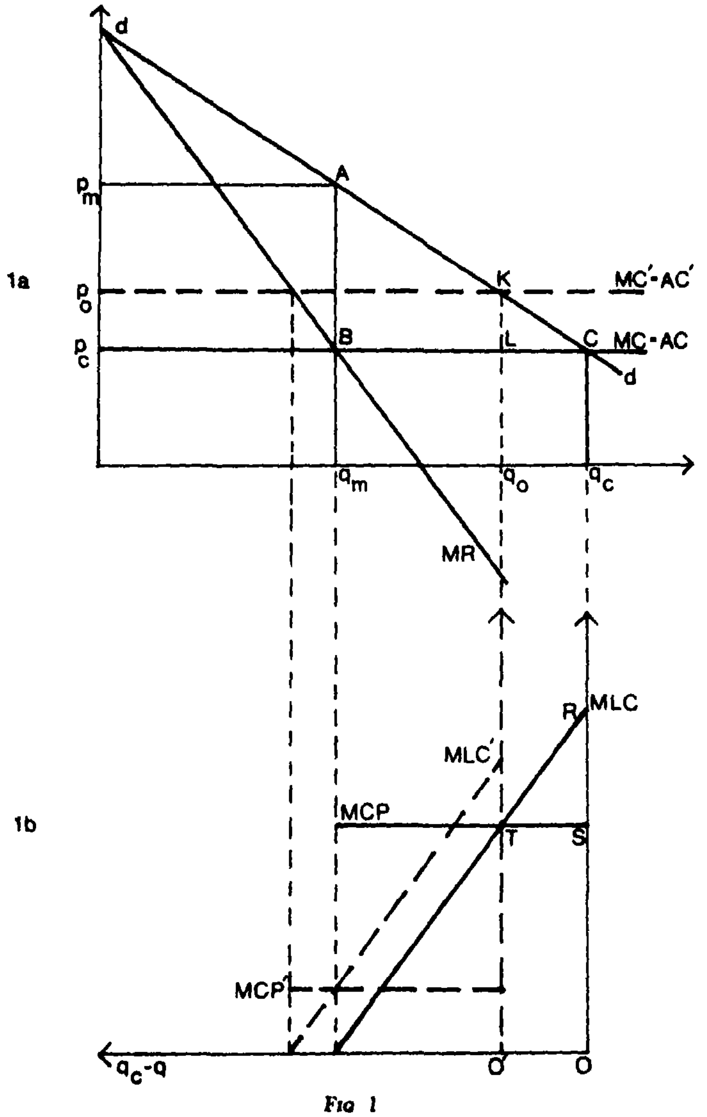


Fig. 1

by the cartel decreases the consumer surplus in the market by p_cp_mAC. Previous calculations of the social costs of monopoly have been based on estimating the "deadweight loss" ACB, as p_cp_mAB represents a transfer to the cartel. Posner's assumptions, however, imply that the whole area p_cp_mAC represents a social loss. He calculates this area by using several earlier estimates of demand elasticities and cartel price increases and concludes that the social costs of monopoly constitute between 5 and 32 percent of the revenues of firms in cartelized industries. Note here that the estimates presented in several of Posner's tables are one-third of the values that follow from his assumptions. This is due to an error in the derivation of an equation.[3]

Posner's calculations are apparently intended to provide an upper limit for the social costs of monopoly, while the deadweight loss *ACB* can be regarded as a lower limit for these costs. However, Posner's assumptions are not sufficient to guarantee that the true social costs of monopoly are less than his estimates. For instance, if a cartel is protected collectively by the judicial system, then costs borne by the judicial system must be added to Posner's estimates (to find an upper limit for social costs). But as Posner himself notes in his introduction: "The estimates are crude; their primary value may simply be to induce skepticism about the existing empirical literature on the social costs of monopoly" (1975, p. 809).[4]

II. The Cartel's Protection Costs

Consider the demand and cost relations in a perfectly competitive market. In figure 1*a*, the market is in equilibrium at point *C*. The sellers in the market wish to form a cartel in order to maximize total profits; this would involve a price increase to the monopoly price p_m. Of course, each seller would in this situation have a private incentive to cheat on the cartel as he would receive a price greater than his marginal cost. If there is such cheating, then the damages to the cartel as a whole will be equal to the area between marginal costs and marginal revenues between the monopoly quantity q_m and the volume of the good actually sold. From this it is possible to derive a marginal damage curve for cheating. This marginal damage of offences against the cartel ("crime") is $MLC \equiv MC - MR$, which is shown in figure 1*b* (where crime protected against is measured on the horizontal axis as $q_c - q$).

If the cartel does nothing to protect itself against crime the competitive solution will obtain. However, assume that the cartel has at its disposal some costly means of protecting itself against violations of its rules. The marginal costs of protection are represented in figure 1*b* as curve *MCP*

---

[3] The calculations are based on equation (10′) (p. 815) which should be $C = 3R_m/2\varepsilon = \frac{3}{2}(1 - k)R_m$, and not $C = R_m/2\varepsilon$.

[4] This crudeness seems to be forgotten, however, when he draws the conclusion "that public regulation is probably a larger source of social costs than private monopoly" (p. 807). The statement is more general than what follows from the rest of the article. Furthermore, property rights are not fixed in Posner's analysis. To "own" a good means, *inter alia*, that one can enjoy the income, or rent, or trade, of the good (see Cheung 1974). This implies that others cannot transfer ownership to themselves without compensating the original owner. Ownership is protected to a large extent collectively, presumably because this is relatively efficient. Thus, when Posner allows costly competition for rents he implicitly drops the assumption that property rights are given, something which draws the foundations out from under many of the (present) results of economic theory. If all rents in an economy are transformed into social costs, for instance, then both "the price mechanism" and "efficiency" become void concepts.

These costs are assumed constant in the figure to maintain consistency with the earlier assumptions made about production costs. It might of course be more reasonable to assume increasing marginal protection costs as more protection will lead to a greater price and thus greater incentives for the cartel's members to cheat, but such an assumption would affect the results only trivially.

Here, there is a protection problem within the cartel of the same type which has been analyzed by Becker (1968) in his essay on crime and punishment. The optimal level of protection (and its corresponding crime rate) is given by point T in figure 1*b* and, tracing upward to figure 1*a*, this involves a sales volume of q_o and a price level of p_o. In this case, the existence of protection costs leads to a lower price and a higher sales volume than in a monopolistic situation without protection costs.[5] The cartel's profits will also be lower than under monopoly, partially as a result of this, and partially since the cartel expends resources for protection, profits will be $p_c p_o KL - O'TSO = TRS$. (I have assumed fixed protection costs to be zero here as a simplification.)

This analysis assumes that the demand and production cost relationships remain fixed during cartelization, something which need not always be the case. The costs of maintaining a cartel might reveal themselves as increases in marginal production costs. For instance, Posner cited the case of an airline cartel which restricts supply and which then faces nonprice competition. Here, a production cost increase due to the cartel arrangement can influence the cartel's price and sales volume (see Posner 1975, pp. 809 ff.). As a somewhat different example, an agreement to divide up a region into several areas, each monopolized by a single firm, might lead both to separate and different demands as well as production costs.

To isolate the effects of a shift in production costs, assume that a cartel's members are willing to accept increased *production* costs ($MC' = AC'$ in fig. 1*a*) in order to decrease the *protection* costs required to limit supply, and that the cartel's price decisions are based on a constant market demand relationship. This change in costs causes MLC to shift to MLC' and MCP to shift to MCP', and it thus leads the cartel's optimal price to be p_m. That is, it is possible that the same price and volume of trade arise with protection costs as without them. If there are lower marginal protection costs than MCP' or greater marginal production costs than MC' then a price which is greater than p_m will be optimal for the cartel (and vice versa if MCP is greater or MC lower than MCP' and MC', respectively).

---

[5] Posner implicitly considers the effect of the cartel's protection costs on its pricing decision when he regards the case when "$MC = MR$" to be a special case where the cartel can extract the whole "supercompetitive" price increase (p. 813).

Of course, a cartel which earns positive profits may face competition from producers of close substitutes Protection against competitors will generally involve shifts in the demand faced by the cartel This situation—which has been discussed in connection with monopolistic competition—will not be taken up here

References

Becker, G. "Crime and Punishment An Economic Approach." *J P E* 76, no 2 (March/April 1968) 169-217

Cheung, S. "A Theory of Price Control " *J Law and Econ* 17 (January 1974) 53-71.

Posner, R. "The Social Costs of Monopoly and Regulation " *J P E* 83, no 4 (August 1975)· 807-27.

Distribution, the Aggregate Consumption Function, and the Level of Economic Development: Some Cross-Country Results

Philip A. Della Valle and Noriyoshi Oguchi

University of Windsor, Ontario

In a recent issue of this *Journal*, Blinder (1975) examined in some detail the potential impact of the distribution of income on the level of consumption in the United States (1947–72). On the basis of his time-series analysis he rejects the Keynesian hypothesis that the equalization of income distribution would increase consumption. In fact, his results suggest either that aggregated consumption is independent of the distribution of income, or that greater income inequality will lead to an increase in aggregate consumption. His empirical tests seem to favor the counterintuitive explanation that increases in inequality result in increases in consumption.

Since we share Blinder's surprise that the relationship between aggregate consumption and income distribution has never been subject to direct empirical tests, we decided to undertake some simple empirical examinations of our own. Our main purpose was to find out whether Blinder's conclusions for United States time-series data would hold true for cross-country consumption patterns. Our analysis of the consumption-distribution relationship leads to a related empirical inquiry, namely, a quantification of the relationship between income distribution and the level of economic development.

I. The Distribution-Consumption Relationship for Developed and Less Developed Countries

Our examination of the cross-country relationship between consumption and distribution began with the estimation of simple consumption functions for 37 developed and less developed countries for the years 1960–71. Using data from the United Nations (1973) we regressed private consump-

[Journal of Political Economy, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved

tion on gross domestic product for all countries in our sample.[1] We realize that the use of GDP instead of the traditional disposable income figure might create some biases in the estimates of average propensity to consume (APC) for each country.[2] However, since we initially wanted to maximize the number of countries in our sample we opted to use the GDP figures Subsequent tests will be performed with disposable income figures which were available for a smaller number of countries.

In our initial tests we decided to use the Gini coefficient as a simple measure of the equality of income distribution The Gini coefficients were taken from Paukert's study (Paukert 1973) and represent estimates for years in the neighborhood of 1965, which is approximately the midpoint of our consumption functions [3]

We examined the impact of income distribution on aggregate consumption by regressing the APC values on the Gini coefficients for the countries in our original sample, and the results are given in equation (1) of table 1 The coefficients and R^2 are significant at the 99 percent level [4] The positive relationship between the APC and the Gini coefficient with cross-country information parallels Blinder's time-series results for the United States.[5] Our initial findings, therefore, support the hypothesis that increases in income inequality (a higher Gini coefficient) result in increases in aggregate consumption Blinder offers some alternative theoretical explanations for his conclusions, including a modification of the permanent income or life-cycle hypothesis, acceptance of the Duesenberry relative income hypothesis, and the existence of changes in demographic and labor force factors which have distorted the measurement of

---

[1] The 37 countries included in our original sample are Nigeria, Morocco, Sierra Leone, Tunisia, Bolivia, Sri Lanka, Pakistan, Fiji, Ivory Coast, Brazil, Ecuador, El Salvador, Peru, Philippines, Colombia, Costa Rica, Jamaica, Chile, Mexico, Panama, Trinidad and Tobago, Venezuela, Greece, Japan, Israel, United Kingdom, the Netherlands, Federal Republic of Germany, France, Finland, Italy, Puerto Rico, Norway, Australia, Denmark, Sweden, and the United States

[2] Those biases might result from such factors as differences in taxation structures, unemployment insurance, pension schemes, and other social welfare programs

[3] We are aware of the empirical problems inherent in using the Gini coefficient as a simple measure of income inequality (Particularly, the problem of different shaped Lorenz curves yielding similar Gini coefficients) Those problems notwithstanding, the Gini coefficient is the most commonly used measurement of inequality, and for that reason we adopt it in our initial tests In Section III we will employ supplementary measures of income distribution to support the empirical results obtained with the Gini coefficient

[4] In table 1 we used a 12-year average value for the APC of each of the 37 countries and regressed those average figures on the independent variables. The resulting equations are cross-country consumption functions which do not rely on any particular assumptions about the relationship between the APC and the marginal propensity to consume (MPC) *within* individual countries.

[5] We also regressed the MPC figures on their corresponding Gini coefficients in order to examine the impact of the level of income distribution on the changes in aggregate consumption. The results were similar but less satisfactory than those of the APC regression.

TABLE 1

CROSS-COUNTRY REGRESSION RESULTS OF APC ON GINI COEFFICIENTS AND INCOME

EQUATION	INDEPENDENT VARIABLES				R^2
	Intercept	G	Y	Y^2	
(1)	0 463	0 47 (3 14)			22
(2)	0 68	0 099 (0 64)	$-0\ 646(10^{-4})$ (−4 08)		48
(3)	0 77	0 012 (0 09)	$-0\ 186(10^{-3})$ (−5 24)	$0\ 5(10^{-7})$ (3 70)	63

NOTE —*t*-values are given in parentheses, income is measured in U S dollars

income inequality in the postwar period Similar interpretations can be applied to our cross-country result

In addition to examining the simple relationship between the APC and the Gini coefficients we included the average level of income in our cross-country regressions This step was taken to eliminate the unduly strong assumption behind equation (1), that is, that the APC does not change as income rises In equations (2) and (3) (table 1) we can see that the linear and nonlinear impact of income (Y) significantly influences the APC The R^2 improves from equation (1) to equation (2) as well as from equation (2) to equation (3), and the *t*-statistics are significant at the 99 percent level for all of the income variables The signs of the coefficients indicate that the APC falls as income rises A result which demonstrates that our cross-country consumption function corresponds in shape to the usual intracountry short-run or cross-sectional function

The inclusion of the level of income, on the other hand, renders the impact of the Gini coefficient insignificant This result, however, does not necessarily imply that income distribution has no influence on aggregate consumption. We can suggest two possible explanations for the lack of significance of the income distribution variable in equations (2) and (3). The first is that there exists a high degree of collinearity between G and Y, and as a result the inclusion of the income variable overrides the influence of G on APC[6] This result would not be surprising given the theoretical and empirical evidence on the relationship between income distribution and the level of per capita income That relationship will be examined further in Section II

The second possibility is that the GDP figures are not a close enough proxy for the more appropriate disposable income figures That is to say,

[6] The regression of G on Y yields the following result

$$G = 50 - 608\,(10^{-4})Y \qquad R^2 = 35$$
$$(4\ 32)$$

consumption is considerably influenced by government participation in the economy through taxes and transfers. Disposable income figures, however, are not available or comparable for the large sample of countries in Section I. In Section III, therefore, we will continue our analysis on the distribution-consumption relationship for 10 OECD countries for which disposable income data were available.

II. The Distribution-Development Relationship

In his now classic study S. Kuznets (1955) observed that income is distributed more unequally in underdeveloped than in developed countries Since then several cross-country studies on the relationship between income distribution and economic development have appeared in the literature Kravis (1960) supported the Kuznets findings (reaffirmed in 1963), while Oshima (1962) classified a country's development into four stages and suggested that income inequality is low at very low levels of economic development As a country advances through the development process, however, the dispersion of incomes increases, peaks, then falls when the country reaches the fully developed stage. It appears that as a country begins to develop and becomes more monetized income inequality increases At higher levels of development, however, progressive taxation and other measures which reflect the welfare ethic redistribute income and hence reduce income inequality.

Recently Paukert (1973) has examined the income distribution-development hypothesis, utilizing and revising the extensive data gathered by I. Adelman and C Taft Morris (1971). Paukert's conclusion on the cross-sectional pattern of income distribution using his more extensive data "supports the hypothesis expressed but not fully tested by Kuznets and Oshima that with economic development income inequality tends to increase, then becomes stable and then decreases" (Paukert 1973, p. 121)

Since we now have access to the large number of observations made available through the Paukert study, it would seem that the above hypothesis could be empirically generalized by regressing a measure of income inequality—the Gini ratio, *G*—on a measure of economic development—gross domestic product per head, *Y*. According to the above hypothesis, the shape of the curve which might best describe the empirical relationship between *G* and *Y* is given by a positively skewed frequency polygon described in equation (4) and in figure 1:

$$G = \alpha e^{b(\ln Y - C)^2}, \tag{4}$$

where *G* is the Gini ratio, *Y* is GDP per capita, and α, b, and *C* are constants.

A positively skewed relationship seems reasonable since the Gini ratio appears to rise quickly with initial increases in the level of *Y* and then

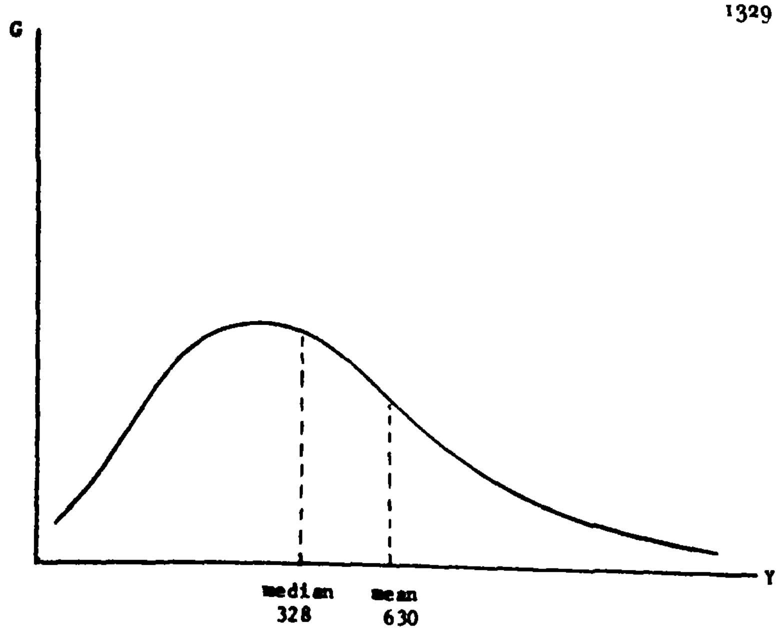


FIG 1 —Hypothesized relationship between the Gini coefficient (G) and income per capita (Y)

falls at a slower rate with further increases in Y In addition, the skewed curve is justified since the mean of Y, 630, is significantly greater than its median, 328 (see fig 1)

Taking logarithms to transform equation (4) into a linear form, we are able to use ordinary least squares (OLS) estimates to solve for the coefficients in that equation The result is given as follows

$$G = 487e^{-0\ 78(\ln Y - 5\ 80)^2}$$

The coefficients are significant at the 99 percent level, and the $R^2 = .21$ Given the large number of cross-sectional observations and a F-value of 7.24, the R^2 is significantly greater than zero at the 99 percent level.[7]

The above results would seem to supply a modest empirical generalization to the well-known hypothesis relating income inequality and economic development. The data used support the hypothesis that as per capita income rises income is initially more unequally distributed.[8] Continued increases in per capita income, however, are associated with a more equal distribution of income More specifically, the relationship between income inequality and economic development is shown to be in the general form of a positively skewed frequency polygon.

---

[7] Some of the Gini coefficients, especially among the less developed countries, did not conform to our a priori judgment about the state of income equality in those countries. As an experiment we dropped those 10 countries from our sample and regressed eq (4) for the remaining 46 countries The result was $G = 509e^{-1\ 12(\ln Y - 5\ 92)^2}$ All the coefficients were highly significant and the R^2 rose considerably, to 51

III. The Distribution-Consumption Relationship for 10 OECD Countries

In this section we will extend the analysis of Section I using data (OECD 1974) for 10 OECD countries for which we have comparable figures on disposable income[9] Such data were not available for all of the 37 countries in our earlier sample[10] It was suggested that taxation and transfer differences in those countries, and the necessitated use of GDP figures, rendered the income-distribution variable insignificant in equations (2) and (3) Although the availability of disposable income figures limits the number of countries in our analysis, we gain the advantage of having a sample of relatively homogeneous, developed countries That is to say, the use of our OECD sample should reduce any extreme biases in our consumption functions which would result from vastly different social welfare programs

In order to override the problem of multicollinearity in the OECD sample we pooled time-series and cross-sectional data for consumption and disposable income for 1961 and 1970 Since year-to-year changes in the Gini coefficient are not likely to be very perceptible we used Paukert's data throughout the time period In addition we used his alternative measures of inequality, including percentile distribution figures and the maximum equalization percentage (MEP)

The regression results of equation (2) using disposable income (Y_d) and the alternative income distribution variables (D) are given in table 2. We also estimated Blinder's dynamic equation,

$$\frac{C}{Y_d} = \alpha + \beta \frac{1}{Y_d} + \gamma D + d \frac{C_{t-1}}{Y_d}, \qquad (5)$$

---

[8] The nonlinear relationship between G and Y for the 56-country sample does not contradict our earlier claim that there was collinearity between those variables in eqq (2) and (3) in table 1 (see also n 6). In the previous consumption-function analysis our sample size consisted of only 37 countries since complete data was not available for many of the less developed countries. The consumption analysis therefore included for the most part countries whose higher income levels placed them on the downward-sloping portion of the curve in figure 1.

[9] We would like to thank an anonymous referee for many useful suggestions which we incorporated into our analysis in this section.

[10] The 10 OECD countries included in our sample are Australia, Denmark, Finland, France, Germany, Italy, the Netherlands, Sweden, United Kingdom, and the United States.

TABLE 2

Regression of APC on Income and Distribution Variables

$$\frac{C}{Y_d} = \alpha + \beta Y_d + \gamma D \text{ (Eq [2])}$$

Variables	Gini	MEP	Percentile of Recipients					
			Below 20	21–40	41–60	61–80	81–95	96–100
Constant	0 929	0 916	0 852	0 852	0 624	0 590	0 732	0 948
Income (Y_d)	0 253(10^{-4}) (3 66)	0 263(10^{-4}) (3 76)	0 324(10^{-4}) (4 73)	0 324(10^{-4}) (4 70)	0 153(10^{-4}) (2 54)	0 176(10^{-4}) (3 38)	0 295(10^{-4}) (4 63)	0 170(10^{-4}) (2 78)
Inequality measure (D)	−0 178 (−2 77)	−0 0021 (−2 35)	−0 91(10^{-3}) (−0 37)	−0 40(10^{-3}) (−0 16)	0 016 (6 90)	0 012 (9 14)	0 005 (3 57)	−0 0037 (−0 652)
R^2	25	23	19	19	46	56	28	44

Note — *t*-values are given in parentheses

TABLE 3

REGRESSION OF APC ON INCOME, DISTRIBUTION VARIABLE, AND LAGGED CONSUMPTION

$$\frac{C}{Y} = \alpha + \beta\left(\frac{1}{Y_d}\right) + \gamma D + d\,\frac{C_{t-1}}{Y_d} \text{ (Eq [5])}$$

VARIABLES	GINI	MEP	PERCENTILE OF RECIPIENTS: Below 20	21–40	41–60	61–80	81–95	96–100
Constant	0.573	0.568	0.534	0.541	0.464	0.472	0.495	0.629
Income $(1/Y_d)$	−30.76	−30.81	−33.93	−34.35	−23.34	−25.89	−33.81	−26.38
	(−3.36)	(−3.36)	(−3.72)	(−3.76)	(−2.72)	(−3.26)	(−3.85)	(−3.09)
Inequality measure (D)	−0.070	−0.0009	−0.0003	−0.0008	0.0079	0.0069	0.0024	−0.0018
	(−1.64)	(−1.59)	(−0.21)	(−0.52)	(4.51)	(5.80)	(2.64)	(−4.19)
Lagged consumption (C_{t-1}/Y_d)	0.447	0.450	0.463	0.464	0.384	0.342	0.437	0.388
	(11.02)	(11.21)	(11.64)	(11.66)	(9.56)	(8.53)	(11.01)	(9.50)
R^2	68	68	67	67	73	.76	69	.72

NOTE.—t-values are given in parentheses.

for the pooled sample of 100 observations Those results are given in table 3.

In contrast to our earlier results the Gini coefficient becomes negative and highly significant in equation (2) In equation (5) it is also negative, but the level of significance falls considerably The MEP and the other percentile measures of income inequality follow a similar pattern with variations in the level of significance The middle and upper percentile measures of income inequality (41–95) have positive coefficients, while the highest (96–100) and two lowest (0–40) have coefficients with negative signs Larger values in the extreme percentile ranges indicate a greater degree of income inequality as do smaller values in the middle ranges. Hence the sign reversal is as expected

All the coefficients for the income and lagged consumption variable in both tables are highly significant, with the expected signs The coefficients of determination improve considerably from equation (2) to equation (5).[11]

The above results would seem to support the Keynesian hypothesis that a more equal income distribution results in higher levels of consumption The reversal of the results for the OECD sample from the larger sample of developed and less developed countries can be attributed to several reasons Two possibilities come to mind The first is that the use of the more acceptable disposable income figure eliminates the taxes and transfer distortions of the larger sample where the available GDP figures were used. The second possibility is that our OECD sample only included countries that were relatively developed Hence those results may only reflect the distribution-consumption relationship as it exists in more developed countries

IV. Summary

In this brief paper we have attempted to cast some empirical light on two distinct but related topics Our primary purpose was to examine the cross-country relationship between income distribution and aggregate consumption for a sample of developed and less developed countries Our first tests, given in Section I, yield results which concur with Blinder's (1975) findings for time series United States data, namely, that increases in income inequality may lead to increases in aggregate consumption, although those statistical results were not strongly significant Second, we quantified the cross-country relationship between income inequality and economic development which was first hypothesized by Kuznets (1955, 1963). Our findings gave some support to the hypothesis that the inequality development relationship traces the general shape of a positively skewed frequency polygon. Third, we extended the distribution-consumption analysis of Section I to include more refined data for 10 OECD countries The results in Section III, conflict with those of the larger sample which included both developed and underdeveloped countries. The tests in the OECD sample give some credence to the traditional Keynesian theory that there is a positive relationship between income equality and the level of consumption.

---

[11] In addition to the above tests we ran some regressions with dummy variables which allowed the MPC to vary from country to country The results of those tests were consistent with those reported in table 3

References

Adelman, I , and Taft Morris, C. "An Anatomy of Patterns of Income Distribution in Developing Nations." Grant AID/csd-2236 Final Report, pt. 3 Evanston, Ill. Northwestern Univ., February 1971.

Blinder, Alan S. "Distribution Effects and the Aggregate Consumption Function." *J.P E* 83, no 3, (June 1975) 447–75.

Kravis, I. "International Differences in the Distribution of Income " *Rev. Econ. and Statis* 42 (November 1960) 408–16

Kuznets, S. "Economic Growth and Income Inequality " *A.E R* 45 (March 1955) · 1–28.

———. "Quantitative Aspects of the Economic Growth of Nations. VIII Distribution of Income by Size " *Econ Development and Cultural Change* 11 (January 1963) 1–80

OECD. *National Accounts Statistics* Paris · OECD, 1974.

Oshima, H. "The International Comparison of Size Distribution of Family Incomes, with Special Reference to Asia " *Rev. Econ. and Statis* 44 (November 1962) 439–45

Paukert, F "Income Distribution at Different Levels of Development A Survey of Evidence " *Internat Labour Rev* 108 (August–September 1973) 97–125.

United Nations *Yearbook of National Account Statistics.* New York. United Nations Publishing Service, 1973.

Firm Demand for Information about Price and Technology

Richard E. Kihlstrom

University of Illinois at Urbana-Champaign

Firms are often required to commit resources to production without complete knowledge of price, technology, or other factors which influence output. Familiar examples occur in agriculture where planting decisions are made before commodity price levels and weather conditions are known. Furthermore, as new crops and farming techniques are introduced, the productivity of these innovations is not always completely understood before they are implemented. The construction industry faces similar problems as do most industries in which there is a significant lag between the time when resources are committed to production and the time when output is forthcoming.

In these situations, firms may be able to acquire, at some cost, information about unknown factors before making a production decision. For example, market surveys and trade publications provide information about price. Similarly, agricultural extension services supply farmers with technological information about new crops and techniques and about soil conditions. The weather bureau is another important source of information to farmers as well as homebuilders. When this kind of information is available, firms must decide on the basis of the costs and benefits involved whether it is worth obtaining. The firm must make an even more complicated demand decision when information is available in increasing degrees of reliability at increasing cost. This situation occurs when the firm has control over the size of a sample or the sophistication of information-gathering techniques.

A rigorous analysis of the determinants of firm spending for information requires a general model of the firm's information demand decision. In the model proposed in this paper, as in Kihlstrom (1974*b*), the statistical concept of sufficiency, introduced by Blackwell (1951, 1953), is used as the basis for a general definition of the "amount of information." This definition identifies the quantity of information with its reliability. Using this measure of information it is possible to formally describe the

This research was supported under National Science Foundation grant SOC 75-05317. The author is grateful for this support.

[*Journal of Political Economy*, 1976, vol. 84, no. 6]

© 1976 by The University of Chicago. All rights reserved.

firm's options when the information is available in continuously varying degrees of reliability, that is, in continuously variable amounts. We are also able to prove that information raises expected profits by permitting the firm to adjust its input demands to levels which are more likely to be appropriate for the true price or technology. Furthermore, we can explicitly compute the expected profits associated with each level of information demand. The resulting expression for expected profits is the basis for a comparative static analysis of firm spending on information In particular, we ask whether information demand, like the demand for conventional inputs, declines when price rises A "revealed preference" proof that information demand is a decreasing function of the price of information is contained in Section II, which follows an introduction to the model. This proof is applicable under very general conditions. In particular, the information available can be information about input or output prices or about technology The firms in question may be either competitors or monopolists.

I. The Model

Let $\pi(z, a)$ be the firm's profit function. This function relates profits to the vector of productive inputs, $z = (z_1, \ldots, z_n) \epsilon R_n^+$,[1] employed by the firm and to the unknown parameter $a \in Q \subseteq R_1$. If technology is uncertain and the firm is competitive,

$$\pi(z, a) = pf(z, a) - w \cdot z, \tag{1}$$

where p is the price of the firm's output, f is the production function, and w is the vector of input prices. If price is uncertain,

$$\pi(z, a) = p(a)f(z) - w \cdot z, \tag{2}$$

where $p(a)$ is the random price variable and $f(z)$ is the production function. If input prices are uncertain, then

$$\pi(z, a) = pf(z) - w(a) \cdot z, \tag{3}$$

where p is the price of output, $w(a)$ is the random vector of input prices, and $f(z)$ is the production function. For a monopolistic firm, equations analogous to (1) and (3) describe the profit function if technology or input prices are unknown. In the monopolistic form of these equations p is replaced by $p[f(z)]$ for equation (3) and by $p[f(z, a)]$ for equation (1) The monopolistic form of (2) is

$$\pi(z, a) = p[f(z), a]f(z) - w \cdot z. \tag{2'}$$

---

[1] R_n is n-dimensional Euclidean space, R_n^+ is the positive orthant of n-dimensional Euclidean space.

We suppose that the firm has expectations about the parameter a which can be represented by an a priori probability distribution with density $g(a)$ If the firm maximizes expected profits, it will choose z to maximize $\int_Q \pi(z, a)g(a)da$, where Q is the set of possible a values The maximizing z is assumed to be unique and is denoted by $z(0)$ [2]

Now suppose that varying amounts of information are available Formally, information is obtained by observing a random variable which is correlated with a If the information is to be of use, the observation must be made before an input vector is committed to production Roughly, random variables which are more closely correlated with a provide better information In the extreme case, a variable which is perfectly correlated with a, for example, a itself, provides perfect information Knowing the value of a variable which is perfectly correlated with a is equivalent to knowing a.

To introduce the possibility of variable amounts of information, we assume that the firm can choose to observe any one of an infinite number of random variables $\{\tilde{x}_\theta\}_{\theta\epsilon(0\ \infty)}$, each of which is correlated with a The random variables $\tilde{x}_\theta$ all take values in a sample space X, and the conditional density of $\tilde{x}_\theta$ given a is $h(x|a, \theta)$

The parameter θ will be interpreted as the "amount of information" conveyed by observing $\tilde{x}_\theta$ To guarantee that information does increase with θ, we make use of the statistical concept of sufficiency in the form introduced by Blackwell (1951, 1953) Thus we assume that $\theta > \theta'$ implies that $\tilde{x}_\theta$ is sufficient, relative to the parameter a, for $\tilde{x}_{\theta'}$ We also assume that when θ exceeds θ', $\tilde{x}_{\theta'}$ is not sufficient for $\tilde{x}_\theta$ Analytically, $\tilde{x}_\theta$ is sufficient for $\tilde{x}_{\theta'}$ when there exists a conditional density function $m(x_{\theta'}|x_\theta)$ such that for all $x_\theta \in X$, $x_{\theta'} \in X$ and all $a \in Q$

$$h(x_{\theta'}|a, \theta') = \int_X m(x_{\theta'}|x_\theta)h(x_\theta|a, \theta)dx_\theta \tag{4}$$

The crucial feature of the conditional density $m(x_{\theta'}|x_\theta)$ is the fact that it is independent of a Thus, once $\tilde{x}_\theta$ is known to equal x_θ, the random variable $\tilde{x}_{\theta'}$ is independent of a, and observations of $\tilde{x}_{\theta'}$ convey no additional information about a. It is easy to prove that when $\tilde{x}_\theta$ is sufficient, relative to a, for $\tilde{x}_{\theta'}$ every user of information about a is at least as well off observing $\tilde{x}_\theta$ before making a decision as he would be if he based his decision on $\tilde{x}_{\theta'}$.[3] Blackwell provides one proof Another proof is given in Marschak and Miyasawa (1968)

---

[2] An example in which this uniqueness condition holds is described in Kihlstrom (1976).

[3] The proof requires a formal definition of an "information user " In essence, a user of information about a is a decision maker whose welfare is determined by the parameter value a as well as by his decision. If an information user knows a, his optimal action will be related to the particular value taken by a

Because better information is usually more expensive we assume that the opportunity to observe $\tilde{x}_\theta$ costs more than the opportunity to observe $\tilde{x}_{\theta'}$ when $\theta > \theta'$. In fact, we assume that $q > 0$ is the price of a "unit of information" and that the cost of θ units of information, obtained by observing $\tilde{x}_\theta$, is $q\theta$.

If $x \in X$, and the firm observes that $\tilde{x}_\theta = x$, the firm will alter its expectations about a via Bayes's rule which asserts that the posterior density of a given that $\tilde{x}_\theta = x$ is

$$\eta(a|x, \theta) = \frac{h(x|a, \theta)g(a)}{v(x|\theta)}$$

where

$$v(x|\theta) = \int_Q h(x|a, \theta)g(a)da \tag{5}$$

The resulting revision in the firm's profit expectations induces an alteration in the input choice Specifically, when $\tilde{x}_\theta$ has been observed to equal x the firm employs an input vector $z(x, \theta)$ which maximizes

$$\int_Q \pi(z, a)\eta(a|x, \theta)da. \tag{6}$$

We will assume that $z(x, \theta)$ is unique for each x and θ.[4] Thus, if the firm observes $\tilde{x}_\theta = x$, its expected profits are

$$\int_Q \pi[z(x, \theta), a]\eta(a|x, \theta)da - q\theta \tag{7}$$

Notice that the cost of observing $\tilde{x}_\theta$ is deducted Also note that (6) and therefore $z(x, \theta)$ are both independent of q

The function $z(x, \theta)$ specifies a contingency plan which describes how the firm adjusts its input decision in response to alternative messages it might receive once it has decided to base its decision on $\tilde{x}_\theta$ But when the firm decides which $\tilde{x}_\theta$ to observe, that is, when it decides how much information to purchase, the firm does not yet know what message the information will bring Thus the decision to purchase information must be based on an expected utility maximization in which the utility associated with each possible message is weighted by the probability of receiving that message The utility of the "message $\tilde{x}_\theta = x$" is given by (7) The probability of the "message $\tilde{x}_\theta = x$" is given by (5) The expected profit associated with observing $\tilde{x}_\theta$ before deciding on an input vector is therefore

$$H(\theta, q) = \int_X \int_Q \pi[z(x, \theta), a]\eta(a|x, \theta)da\,v(x|\theta)dx - q\theta, \tag{8}$$

---

[4] This uniqueness condition holds in the example described in Kihlstrom (1976)

when $\theta \in (0, \infty)$ If $\theta = 0$,

$$H(0, q) = \max_{z} \int_{Q} \pi(z, a) g(a) da = \int_{Q} \pi[z(0), a] g(a) da$$ [5]

When the firm makes an information demand decision it does so by choosing θ to maximize $H(\theta, q)$ For each q, the maximizing value of θ is assumed to be unique [6] It is denoted by $\theta(q)$

In the analysis which follows we shall assume that information markets are competitive As usual, this implies that in choosing θ to maximize H the firm treats q as a parameter outside of its control Clearly the level of demand for information will be determined by q as well as by the parameters of the profit function and of the a prion density function g When input markets (other than the information market) are competitive, input prices will be among these parameters If the output market is competitive then output prices will also be included as parameters In the next section, we consider the model in its general form and investigate the effect of information price changes on information demand

II. A Proof That Information Is a "Non-Giffen Input"

The classical theory of the competitive firm asserts that demand for an input never increases when the price of that input rises This result can be proved in several ways, a proof using calculus is given in Samuelson (1947) But it is also possible to give a proof based on a "revealed-preference" argument In the theorem which follows, the revealed-preference approach is applied to demonstrate that, when firms maximize expected profits, the firms' information demand $\theta(q)$ will not increase if the price of information rises This result holds for input and output price information as well as for technological information It also holds for monopolists who buy information in competitive markets In the monopolistic case the information may be information about technology, input prices, or the demand curve

Theorem.—Assume that an expected profit-maximizing information demand $\theta(q)$ exists for each possible information price level q Then information demand $\theta(q)$ decreases or remains constant when the price of information rises

---

[5] We assume throughout that

$$\lim_{\theta \to 0} H(\theta, q) = H(0, q)$$

The example in Kihlstrom (1976) does result in a $H(\theta, q)$ function which is continuous at zero.

[6] This uniqueness condition holds in the example described in Kihlstrom (1976)

Proof.[7]—Let $q_1 > q_2$. By definition of $\theta(q_1)$ and $\theta(q_2)$,

$$H[\theta(q_1), q_1] \geq H[\theta(q_2), q_1] \tag{9}$$

and

$$H[\theta(q_2), q_2] \geq H[\theta(q_1), q_2]. \tag{10}$$

Using the fact that $H(\theta, q) = H(\theta, 0) - q\theta$, equation (9) can be rearranged to obtain

$$H[\theta(q_1), 0] - H[\theta(q_2), 0] - q_1[\theta(q_1) - \theta(q_2)] \geq 0. \tag{11}$$

Similarly, equation (10) reduces to

$$H[\theta(q_2), 0] - H[\theta(q_1), 0] - q_2[\theta(q_2) - \theta(q_1)] \geq 0. \tag{12}$$

Adding the left-hand sides of (11) and (12) yields $(q_2 - q_1)[\theta(q_1) - \theta(q_2)] \geq 0$. Since $q_2 < q_1$, $\theta(q_2)$ must therefore exceed or equal $\theta(q_1)$

Summary

The model employed in this analysis assumes that information about a parameter, say price, is provided by observations of random variables correlated with price. It is also assumed that the firm is able to observe one of many different random variables, each of which is correlated with price The statistical concept of sufficiency is used to define analytically the amount of price information conveyed by each possible observation. For the purpose of analyzing, in general, the determinants of information demand, this definition is the crucial feature of the model. Using this definition, information increases can be shown to raise the level of expected profits, but they also are assumed to be expensive As with other commodities, the optimal level of information demand occurs where the marginal benefits of more information are exactly offset by the marginal costs The analysis in the present paper uses the model in its general form to demonstrate that the optimal level of information demand declines as the price of information increases.

It is also possible to study other determinants of information demand with this same model. This is done, for a special case, in Kihlstrom (1976) The special case treated there occurs when the probability distributions are normal. In the normal case, the amount of information can be measured by the inverse of the variance of the observation. It can also be interpreted as "sample size." For further remarks on these interpretations the reader is referred to Grossman, Kihlstrom, and Mirman (1975) and Kihlstrom (1974a, 1976).

---

[7] I am grateful to Ed Foster for suggesting this simple proof. Earlier versions of this paper contained a much longer proof based on the same idea. I am also grateful to the editors of the *Journal of Political Economy* for allowing me to substitute the simpler proof at a very late stage in the publication process.

References

Blackwell, David. "The Comparison of Experiments." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* Berkeley Univ. California Press, 1951

———. "Equivalent Comparison of Experiments" *Ann Math Statis* 24 (June 1953): 265–72.

Grossman, Sandford, Kihlstrom, R E, and Mirman, L J "A Bayesian Approach to the Production of Information and Learning by Doing" Technical Report no. 185, Economics Series, Stanford Univ, Inst. Math Studies Soc Sci, October 1975.

Kihlstrom, R E. "A Bayesian Model of Demand for Information about Product Quality." *Internat Econ. Rev* 15 (February 1974)· 99–118. (*a*)

———. "A General Theory of Demand for Information about Product Quality" *J. Econ. Theory* 8 (August 1974) 413–40 (*b*)

———. "Technological and Market Determinants of Firm Information Demand An Example." Unpublished paper, Univ Illinois at Urbana-Champaign, January 1976.

Marschak, Jacob, and Miyasawa, Koichi "Economic Comparability of Information Systems" *Internat. Econ Rev* 9 (June 1968) 137–74

Samuelson, Paul *The Foundations of Economic Analysis* Cambridge, Mass · Harvard Univ. Press, 1947

The Desirability of Cheating Incentives and Randomness in the Optimal Income Tax

Laurence Weiss

Harvard University

A common feature of income tax plans is the opportunity to evade taxes illegally To make this unattractive, governments audit a random sample and punish agents who cheat, usually with fines The decision to cheat, from the agent's viewpoint, is analogous to the purchase of a risky asset The availability of risky financial assets will generally affect the relative attractiveness between labor and leisure It follows that the creation of incentives to cheat will alter the amount of labor agents wish to supply for any given wage and tax rate

It is well known that an income tax distorts the marginal incentives and lowers the supply of labor below that which agents would choose if, instead, a lump-sum tax which yielded the same revenue were levied This paper will explore the possibility that incentives to cheat might offset this undesirable substitution effect and improve resource allocation sufficiently to compensate agents for their usual aversion to risk

As in Mirrlees's (1971) important article on the optimal redistributive income tax, an atemporal, artisan economy (no firms, no capital, constant marginal product, labor-leisure trade-off) is analyzed The government employs an income tax to raise an exogenous amount of revenue For expositional simplicity, only linear tax schedules are examined, but the restriction is not crucial. The real-world consideration that much effort is wasted by devising and thwarting evasion schemes is assumed away by postulating that the acts of evasion and detection are, in themselves, costless. Lump-sum taxation is ruled inadmissable, so that the problem is inherently in the nature of the second best

A simple model of tax evasion formulated by Allingham and Sandmo (1972), Srinivasan (1973), and others is extended to allow for variable labor supply. It is observed that, starting from a position in which no one

[*] Financial support from a National Science Foundation graduate fellowship and NSF grant no. SOC 73-09142 is gratefully acknowledged I would like to thank K Arrow, S. Shavell, and members of the Institute for Mathematical Studies in the Social Sciences at Stanford University for helpful discussions An anonymous referee provided valuable suggestions, and Mike Bazdarich corrected several errors in an earlier draft

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved.

cheats, it is to the advantage of all agents to have incentives to cheat, while these incentives might be provided without any loss of (expected) tax revenues. The optimal income tax may, therefore, provide incentives to cheat. This leads to the analogous and somewhat paradoxical idea that, to raise a given amount of tax revenue, all agents might prefer a random tax rate rather than a certain one which produced the same revenue, even though they are risk averse in the usual sense (have concave utility functions). The possibility of this is examined when the realization is announced both before and after labor supply is chosen The latter scheme will be shown to be desirable under the same conditions which imply that the provision of incentives to cheat is desirable.

This suggests that it might be desirable (in the sense of increasing all agents' expected utility) if the taxing authority treats identical individuals differently Whether a social planner should be governed solely by agents' (ex ante) expected-utility calculations is a much discussed and open question, but most economists would probably accept individuals' assessments of their own welfare when these are made with the "correct" probabilities

Cheating Strategy

All agents are assumed to have preferences over wealth (consumption), denoted by W, and leisure (L) that may be represented by a von Neumann-Morgenstern (cardinal) utility function Following Allingham and Sandmo (1972), the penalty assessed on fraudulent returns is proportional to the amount of the deficiency The government announces that it will audit $(1 - p)$ of the returns. Those audited and found deficient will be required to pay a penalty of q times the income understated, in addition to the normal tax of t

Normalizing the wage rate at one, all agents choose hours worked (h) and income understated (C) so as to maximize

$$F(h, C) = E[U(W, L)]$$
$$= pU[h(1 - t) + tC, -h] + (1 - p)U[h(1 - t) - qC, -h] \tag{1}$$

First-order conditions for an interior maximum imply

$$\frac{\partial F}{\partial h} = p[U_1^+(1 - t) - U_2^+] + (1 - p)[U_1^-(1 - t) - U_2^-] = 0 \tag{2}$$

and

$$\frac{\partial F}{\partial C} = ptU_1^+ - q(1 - p)U_1^- = 0. \tag{3}$$

(Superscript $^+$ means the function is evaluated at $h[1 - t] + tC$, income if not audited; superscript $^-$ means the function is evaluated at

$h[1 - t] - qC$, net income if audited) Expected tax revenues are $E(T) = h^*t - [pt - (1 - p)q]C^*$, where h^* and C^* are chosen to satisfy the first-order conditions Sufficiency is established by observing that $F(h, C)$ is concave, being the average of concave functions of linear transformations of h and C

Suppose the audit rate is set just high enough so that no agent has an incentive to cheat By equation (3), this implies $pt - (1 - p)q = 0$ (so that $p = q/[t + q]$), and the expected gain from cheating is zero To examine the change of h^* and C^* in response to enforcement changes, equations (2) and (3) are totally differentiated with respect to p in the usual Hicksian manner Evaluated at $p = p^* = q/(t + q)$,

$$\left.\frac{\partial h^*}{\partial p}\right|_{p=p^*} = 0,$$

$$\left.\frac{\partial C^*}{\partial P}\right|_{p=p^*} = -\frac{U_1}{U_{11}}\frac{t+q}{tq},$$

$$\left.\frac{\partial E(T)}{\partial p}\right|_{p=p^*} = t\frac{\partial h^*}{\partial p} - (t + q)C^* - [pt - (1 - p)q]\frac{\partial C^*}{\partial p} = 0$$

(see Appendix, item 1)

Hence, if p is raised slightly, all people cheat inversely proportionally to their coefficient of absolute risk aversion It is necessary to examine the second derivatives to ascertain the changes in labor supply and expected taxes.

At $p = p^*$,

$$\frac{\partial^2 h^*}{\partial p^2} = \frac{(t + q)^2}{tq}\left\{\frac{U_1[U_{111}(1 - t) - U_{121}]}{U_{11}^2} - 2(1 - t) + 2\frac{U_{12}}{U_{11}}\right\}, \quad (4)$$

where $\Delta = -U_1/[U_{11}(1 - t)^2 - 2U_{12}(1 - t) + U_{22}]$ = minus the substitution effect of the change in the tax rate (>0)

In general, the sign of the expression inside the parentheses is ambiguous, reflecting opposing forces Cheating has a positive expectation, and everyone understates his true income This raises expected wealth for a given effort and tends to lower the marginal utility of wealth and reduce the optimal hours worked Cheating makes final wealth random This uncertainty will lower the inducement to work if $U_1(1 - t) - U_2$ is concave in wealth and will raise optimal labor supply if it is convex In the case of utility functions that are separable in income and leisure, a positive sign of equation (4) is assured by assuming that absolute risk aversion decreases rapidly with wealth, since $d[-(U_{11}/U_1)] = -(U_{111}U_1 - U_{11}^2)/U_1^2 = -(U_{11}^2/U_1^2)\{[(U_{111}U_1)/(U_{11}^2] - 1\}$ If the expression above is less than $-(U_{11}^2/U_1^2)$, labor supply will increase with the opportunities to evade taxes, if the utility function is separable This problem is similar to asking how variability of the interest rate affects the supply of savings (for a good discussion of this and similar problems, see Rothschild and Stiglitz [1971] and Levhari [1973]).

The change in expected tax receipts is

$$\frac{\partial^2 E(T)}{\partial p^2} = t\frac{\partial^2 h^*}{\partial p^2} - 2(t + q)\frac{\partial C^*}{\partial p} - [pt - (1 - p)q]\frac{\partial^2 C^*}{\partial p^2}, \quad (5)$$

which, evaluated at $p = p^*$, becomes

$$\left.\frac{\partial^2 E(T)}{\partial p^2}\right|_{p=p^*} = \frac{t(t + q)^2\Delta}{tq}\left\{\frac{U_1}{U_{11}^2}[U_{111}(1 - t) - U_{121}] - 2(1 - t) + 2\frac{U_1}{t\Delta U_{11}} + 2\frac{U_{12}}{U_{11}}\right\} \quad (5')$$

This is positive for

$$\frac{U_1}{U_{11}^2}[U_{111}(1 - t) - U_{121}] > 2(1 - t) - 2\frac{U_{12}}{U_{11}} - 2\frac{U_1}{t\Delta U_{11}} \quad (6)$$

The possibility that tax receipts might actually rise with reduced enforcement establishes that a well-designed tax system might provide cheating incentives, since each taxpayer could remain as well off as in the full-audit situation by reporting his entire income. Even if tax receipts decline, it might still be possible to increase the tax rate to compensate the government and leave all agents better off than in the initial situation If t is increased and agents respond accordingly,

$$\frac{\partial T}{\partial t} = h + t\frac{\partial h}{\partial t} = h - t\Delta\left\{1 + \frac{h[U_{11}(1 - t) - U_{12}]}{U_1}\right\}, \quad (7)$$

which will be assumed positive, since otherwise everyone can be made better off simply by lowering t, the marginal tax rate The utility loss to agents is the same whether or not they adjust their labor supply:

$$\frac{\partial U}{\partial t} = -hU_1, \quad (8)$$

so that the trade-off using a straight tax policy is given by (dividing eq. [8] by eq. [7]),

$$\frac{\partial U}{\partial T} = -U_1\bigg/\left\{1 - t\Delta\left[\frac{1}{h} + \frac{U_{11}(1 - t) - U_{12}}{U_1}\right]\right\}. \quad (9)$$

The gain in terms of utility, by providing incentives for gainful tax evasion, is

$$\left.\frac{\partial^2 E(U)}{\partial p^2}\right|_{p=p^*} = -\frac{U_1^2}{U_{11}}\frac{(t + q)^2}{tq}. \quad (10)$$

Dividing the equation above by equation (5), the trade-off by lowering the audit rate is

$$\frac{\partial E(U)}{\partial E(T)} = \frac{-U_1}{(t\Delta/U_{11})[U_{111}(1-t) - U_{121}] - [(2\Delta U_{11})/U_1]t(1-t) + 2[(t\Delta U_{12})/U_1] + 2} \tag{11}$$

Thus the change in utility by raising t and p to leave (expected) taxes unchanged will be positive if and only if equation (9) − equation (11) > 0 or

$$\frac{U_1}{U_{11}^2}[U_{111}(1-t) - U_{121}] > -\frac{U_1 + hU_{12}}{hU_{11}} - \frac{U_1}{t\Delta U_{11}} + (1+t) \tag{12}$$

Random Tax Rates

The preceding analysis suggests that cheating may give rise to a welfare gain by making final wealth variable. This variability may lessen the "dead weight" loss of using an income tax sufficiently to compensate agents for their usual adversity to risk. It seems natural to ask if random tax rates might serve the same purpose.

Suppose agents are told that the tax rate will be a random variable with the same mean as before. The agent must choose pretax income before the realization is announced. Let h^* be the optimal hours worked when the rate is fixed at the expected rate. Thus, h^* satisfied,

$$\frac{\partial U}{\partial h} = (1-t)U_1 - U_2 = 0 \tag{13}$$

For "small" increases in the variance at t, when the variance is zero, hours worked will increase if $(\partial U)/(\partial h)$ is a convex function of t evaluated at h^*:

$$\frac{\partial^2}{\partial t^2}\left(\frac{\partial U}{\partial h}\bigg|_{h=h^*}\right) = h\{2U_{11} + h[U_{111}(1-t) - U_{121}]\} \tag{14}$$

Letting h adjust to restore the (expected) optimality condition of equation (13),

$$\frac{\partial h^*}{\partial \operatorname{var} t} = \frac{h\Delta}{2U_1}\{2U_{11} + h[U_{111}(1-t) - U_{121}]\} \tag{15}$$

Since, by assumption, t has the same mean, the change in expected revenue is

$$\frac{\partial E(T)}{\partial \operatorname{var} t} = t\frac{\partial h^*}{\partial \operatorname{var} t} = \frac{t\Delta h}{2U_1}\{2U_{11} + h[U_{111}(1-t) - U_{121}]\} \tag{16}$$

Similarly, the loss to agents by randomizing is the same whether they adjust optimally or not and is given by the second term of the Taylor expansion of utility with respect to the tax rate about the mean rate:

$$\frac{\partial E(U)}{\partial \operatorname{var} t} = \tfrac{1}{2}h^2 U_{11} \tag{17}$$

So the trade-off between expected utility and expected taxes by randomizing is given by (dividing eq [17] by eq. [16])

$$\frac{\partial E(U)}{\partial E(T)} = -U_1 \Big/ \left\{ t\Delta \left[-\frac{2}{h} - \frac{U_{111}(1-t) - U_{121}}{U_{11}} \right] \right\}. \tag{18}$$

Thus there exists some random rate with a lower mean that is preferred by all agents and yields the same revenue as a certain rate t if $-$ equation (18) + equation (9) > 0 or

$$\frac{U_1}{U_{11}^2}[U_{111}(1-t) - U_{121}] > -\frac{U_1 + hU_{12}}{hU_{11}} - \frac{U_1}{t\Delta U_{11}} + (1-t), \tag{19}$$

which is identical to the condition in equation (12) That is, the possible desirability of permitting incentives to cheat is implied by, and implies, the desirability of using a random rate in preference to a certain rate yielding the same revenue.

Two Examples

If agents' preferences are representable by a constant elasticity of substitution function of the form $U = (\delta W^{-\rho} + (1-\delta)L^{-\rho})^{-1/\rho}$, then expected tax revenues rise with incentives to cheat if (eq. [6]) $t > 2/(1+\sigma)$, where $\sigma = 1/(1+\rho)$, the elasticity of substitution between income and leisure Furthermore, there exists some combination of raising t and p or using some random rate in preference to a certain rate of (eqq [12] and [19])

$$\frac{-\rho}{\rho+1} > \frac{1-t}{t} + \left(\frac{\delta}{1-\delta}\right)^{-1/(1+\rho)} \left(\frac{1}{t}\right)(1-t)^{(2\rho+1)/(\rho+1)}.$$

Particularly, it is necessary that ρ be less than zero, if the elasticity of substitution be greater than one for the schemes to be possibly useful. A Cobb-Douglas function ($\rho = 0$) precludes the desirability of random taxes. This might be easily observed, since hours worked is independent of the tax rate. Any random rate will generate the same labor supply and produce the same expected revenue as a certain rate at the mean, while being deleterious to all agents. This holds whether the realization is announced before or after labor supply is chosen.

If agents' preferences are representable by a utility function of the form $U(W) + L$, where $U(\ )$ is concave and increasing, cheating incentives raise taxes if $(U_1 U_{111})/U_{11}^2 > 2/t$. There exists some cheating strategy, and there also exists some random tax rate, both of which are Pareto preferable to an ordinary linear tax schedule if $(U_1 U_{111})/U_{11}^2 > -(U_1/WU_{11}) + 1/t$. If the condition above is satisfied, there also exists a gamble involving t, the realization of which is antecedent to the selection of pretax earnings, which is also strictly preferable to a fixed t (see [3] in Appendix). For functions of the form $[1/(1-b)](a+W)^{1-b} + L$, $(b > 0)$, cheating increases taxes if $t > 2b/(b+1)$, and all forms of randomization and cheating are potentially useful if $t > b$.

Conclusion

The results suggest that cheating and random taxes are socially useful when U_{111} is large and positive. It is usually assumed that "rational" behavior is characterized by U_1 positive and U_{11} negative. What bounds are reasonable on U_{111}? It has been suggested (Arrow 1971) that absolute risk aversion $(-U_{11}/U_1)$ is nonincreasing and relative risk aversion $(-WU_{11}/U_1)$ nondecreasing with wealth. These conditions imply, respectively, that $1 \leq (U_{111}U_1)/U_{11}^2 \leq (U_1/WU_{11}) + 1$. The right-hand inequality implies that, for arbitrary utility functions that are separable in consumption and leisure, both incentives to cheat and random tax rates are undesirable. That is, if relative risk aversion is a nondeclining function of wealth, then there exists a certain linear tax which yields the same revenue that all agents find preferable to the stochastic schemes. However, if the utility function is nonseparable, the idea of risk aversion becomes less usable and clear. If a constant elasticity of substitution utility function is accepted as a reasonable portrayal of agents' preferences, then the possibility of socially desirable incentives to cheat must also be accepted.

It may seem strange that the introduction of an artificial and unnecessary uncertainty may improve the welfare of a society of risk-averse individuals. The possibility of this is greatest when agents have a strong preference for "skewness" in the distribution of wealth (which is a cardinal property) and when the welfare loss imposed through an income tax is large (which is an ordinal property). The result is less surprising when the second-best nature of the problem is stressed. If lump-sum taxation were allowed (or, equivalently, labor supply were unalterable), the schemes would be useless.

The exclusion of partial enforcement and, more important, of randomization in both Mirrlees's (1971) article on the optimal redistributive tax (so as to maximize the sum of utilities) and Phelps's (1973) problem of maximizing total tax receipts (in accord with Rawlsian justice principles) was nontrivial and without explicit justification. The ethical desirability of treating identical individuals identically (in an ex post sense) must rest on factors other than utilitarianism in a world of the second best

Appendix

1 $(\partial h)/(\partial p)$ and $(\partial C)/(\partial p)$ are defined implicitly by

$$\begin{bmatrix} \dfrac{\partial^2 F}{\partial h^2} & \dfrac{\partial^2 F}{\partial h\,\partial C} \\ \dfrac{\partial^2 F}{\partial h\,\partial C} & \dfrac{\partial^2 F}{\partial C^2} \end{bmatrix} \begin{bmatrix} \dfrac{\partial h}{\partial p} \\ \dfrac{\partial C}{\partial p} \end{bmatrix} + \begin{bmatrix} \dfrac{\partial^2 F}{\partial h\,\partial p} \\ \dfrac{\partial^2 F}{\partial C\,\partial p} \end{bmatrix} = 0,$$

which implies that

$$\begin{bmatrix} \dfrac{\partial h}{\partial p} \\ \\ \dfrac{\partial C}{\partial p} \end{bmatrix} = -\begin{bmatrix} \begin{array}{l} p[(1-t)^2 U^+_{11} \\ \quad - 2(1-t)U^+_{12} + U^+_{22}] \\ \quad + (1-p)[(1-t)^2 U^-_{11} \\ \quad - 2(1-t)U^-_{12} + U^-_{22}] \end{array} & \begin{array}{r} pt[(1-t)U^+_{11} - U^+_{12}] \\ -(1-p)q[(1-t)U^-_{11} \\ - U^-_{12}] \end{array} \\ \begin{array}{l} pt[(1-t)U^+_{11} - U^+_{12} \\ \quad -(1-p)q[(1-t)U^-_{11} \\ \qquad\qquad - U_{12}] \end{array} & pt^2 U^+_{11} + (1-p)q^2 U^-_{11} \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} [(1-t)U^+_1 - U^+_2] - [(1-t)U^-_1 - U^-_2] \\ tU^+_1 + qU^-_1 \end{bmatrix}$$

2 Let $\tilde{t}$ be the random tax rate with variance σ_t^2 Let h^* be the optimal hours worked when $\tilde{t}$ is fixed at its expectation $\bar{t}$ Thus $U_h(h^*, \bar{t}) = 0$ We wish to find dh such that $E[U_h(h^* + dh, \tilde{t})] = 0$ Assuming $U(\ )$ has continuous fourth partial derivatives bounded over the range of t, h under discussion and t is confined to the unit interval,[1] $E[U_h(h, t)] = E\{U_h(h^*, \bar{t}) + (h - h^*)U_{hh} + (\tilde{t} - \bar{t})U_{ht} + \frac{1}{2}(h - h^*)^2 U_{hhh} + \frac{1}{2}(\tilde{t} - \bar{t})^2 U_{htt} + (h - h^*)(\tilde{t} - \bar{t})U_{hth} + 0[(h - h^*)^3] + 0[(h - h^*)^2(\tilde{t} - \bar{t})] + 0[(h - h^*)(\tilde{t} - \bar{t})^2] + 0[(\tilde{t} - \bar{t})^3]\}$, which equals $(h - h^*)U_{hh} + \frac{1}{2}\sigma_t^2 U_{htt} + o(h - h^*) + o(\sigma_t^2)$, as σ_t^2 (and hence, by continuity, $[h - h^*]$) becomes small Thus $(h - h^*) \simeq [(\frac{1}{2}U_{htt})/U_{hh}]\sigma_t^2$ or, at $\sigma_t^2 = 0$, $(\partial h)/(\partial \sigma_t^2) = (\frac{1}{2}U_{htt})/U_{hh}$

3 We seek to answer the question When will it be possible for the government and the individual to agree to a gamble involving t in preference to a fixed t, the realization prior to the agents' selection of labor supply? The general case proved intractible, so attention is confined to utility functions of the form $U(W) + L$

First assume the gamble has the same expected value of t. Differentiating both utility and taxes twice with respect to t, letting h adjust optimally,

$$\frac{\partial E(U)}{\partial \text{ var } t} = -\frac{1}{2}\frac{U_1^2}{(1-t)^2 U_{11}}\left(1 + \frac{2WU_{11}}{U_1}\right),$$

$$\frac{\partial E(T)}{\partial \text{ var } t} = -\frac{1}{2}\frac{U_1}{(1-t)^2 U_{11}}\left[t\left(\frac{U_{111}U_1}{U_{11}} - 2\right) - \frac{2WU_{11}}{U_1} - 2\right],$$

---

[1] In expansions, "0( )" means "terms of order at most," and "o( )" means "terms of smaller order than."

so that, if

$$\frac{U_{111}U_1}{U_{11}^2} > \frac{2}{t}\left(\frac{WU_{11}}{-U_1} + 1 + t\right) \quad \text{and} \quad -\frac{WU_{11}}{U_1} < \frac{1}{2},$$

a random rate is preferred by both taxpayers and the government to a certain rate at the mean

Now suppose $-(WU_{11}/U_{11}) > \frac{1}{2}$, that ex ante randomization hurts the agent but continues to increase expected tax revenues This loss in utility is compared with that incurred by raising the rate to yield the same revenue

$$\frac{\partial E(U)}{\partial E(T)_{\text{randomization}}}$$

$$= \left[U_1(1-t)\left(1 + \frac{2WU_{11}}{U_1}\right)\right] \bigg/ \left[t\frac{U_{111}U_1}{U_{11}^2} - \frac{WU_{11}}{U_1} - 2(1+t)\right],$$

$$\frac{\partial U}{\partial T_{\text{tax rate}}} = -[U_1(1-t)] \bigg/ \left(1 + \frac{tU_1}{WU_{11}}\right)$$

Randomization is superior to increasing taxes if

$$1 + t\frac{U_1}{WU_{11}} < \left[t\left(\frac{U_{111}U_1}{U_{11}} - 2\right) - \frac{WU_{11}}{U_1} - 2\right] \bigg/ \left[-\left(1 + \frac{2WU_{11}}{U_{11}}\right)\right]$$

Since the denominator of the right-hand expression is assumed positive, this is equivalent to

$$\left(1 + \frac{tU_1}{WU_{11}}\right)\left(1 + \frac{2WU_{11}}{U_1}\right) < t\left(\frac{U_{111}U_1}{U_{11}^2} - 2\right) - \frac{2WU_{11}}{U_1} - 2 \qquad \text{(A1)}$$

Now suppose $[\partial E(U)]/(\partial \operatorname{var} t) > 0$, but $[\partial E(T)]/(\partial \operatorname{var} t) < 0$ Comparing randomization with a lowering in t, it is found that randomization is a superior method of lessening the tax burden if

$$1 + \frac{tU_1}{WU_{11}} > \left[t\left(\frac{U_{111}U_1}{U_{11}^2} - 2\right) - \frac{WU_{11}}{U_1} - 2\right] \bigg/ -\left(1 + \frac{2WU_{11}}{U_1}\right), \qquad \text{(A2)}$$

which is equivalent to condition (A1) above as the denominator of the right-hand expression is assumed negative.

Conditions (A1) and (A2) are both equivalent to $(U_{111}U_1)/U_{11}^2 > -(U_1/WU_{11}) + (1/t)$, which is identical to the condition for cheating incentives and randomization, the realization of which is announced after the selection of labor supply, to be preferred to an ordinary linear tax, when utility functions are of the form $U(W) + L$.

References

Allingham, M. G., and Sandmo, A "Income Tax Evasion A Theoretical Analysis." *J. Public Econ.* 1 (November 1972) 323–34

Arrow, K. J. *Essays in the Theory of Risk Bearing* New York Markham, 1971

Levhari, D. "Optimal Savings and Portfolio Choice under Uncertainty." In *Mathematical Methods in Investment and Finance*, edited by V. P. Szego and K. Shell. Amsterdam: North-Holland, 1973.

Mirrlees, J. A. "An Exploration in the Theory of Optimum Income Taxation." *Rev. Econ. Studies* 38 (April 1971): 175–208.

Phelps, E. S. "Taxation of Wage Income for Economic Justice." *Q.J.E.* 87 (August 1973): 331–54.

Rothschild, M., and Stiglitz, J. "Increasing Risk. II. Its Economic Consequences." *J. Econ. Theory* 3 (March 1971): 66–85.

Srinivasan, T. N. "Tax Evasion: A Model." *J. Public Econ.* 2 (July 1973): 257–65

On the Specification of the Demand for Money: The Real Rate of Return versus the Rate of Inflation

Benjamin Eden

The Hebrew University, Jerusalem

I

The demand for money is sometimes specified as a function of the anticipated rate of inflation and sometimes as a function of the anticipated real rate of return on money.[1] In the absence of uncertainty, individuals hold point anticipations and it does not matter which specification is used, since the real rate of return and the rate of inflation are monotonically related.

When consumers are uncertain about future prices, it makes a difference whether anticipations with respect to the mathematical expectations of the real rate of return on money or anticipations with respect to the expected rate of inflation are used to explain changes in the demand for money. The difference arises from the fact that the rate of inflation is not a linear function of the real rate of return. In this note, we provide evidence that supports the specification of the demand for money as a function of anticipations with respect to the expected real rate of return on money, over the usual specification which uses the expected rate of inflation.

II

Consider a consumer who lives in a world in which money is the only available asset and there is an aggregate consumption good. The consumer anticipates *n* possible states of the world (i.e., *n* possible evolutions of the world) which are characterized by the consumer's future income

This note is based on the first chapter of my Ph.D. dissertation which was submitted to the University of Chicago. I would like to thank the committee members W. A. Brock, R. J. Barro, S. Fischer, and M. Friedman. I also benefited from comments provided by J. Frenkel and two anonymous referees.

[1] See, e.g., Friedman (1956) for the first specification. The second specification stems from the portfolio approach to the demand for money (see Tobin 1958).

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved.

and the future price level. He uses these anticipations to plan his consumption and holdings of real balances for two periods.[2] Let

C_0 = level of consumption in period 0;
C_1^i = level of consumption in period 1 in state of the world i $(i = 1, \ldots, n)$;
M_0 = quantity of nominal balances in period 0;
P_0 = the price level in period 0 (the price of consumption in terms of money);
$m_0 = M_0/P_0$ = quantity of real balances in period 0,
M_1^i = quantity of nominal balances in period 1 in state of the world i $(i = 1, \ldots, n)$;
P_1^i = price level in period 1 in state of the world i $(i = 1, \ldots, n)$;
$m_1^i = M_1^i/P_1^i$ = quantity of real balances in period 1 in state of the world i $(i = 1, \ldots, n)$;
W_0 = initial real wealth;
I_1^i = nominal income in period 1 in state of the world i $(i = 1, \ldots, n)$,
$W_1^i = I_1^i/P_1^i + M_0/P_1^i$ = real wealth in period 1 in state of the world i $(i = 1, \ldots, n)$,
q_i = the consumer's subjective probability that state of the world i will occur $(i = 1, \ldots, n)$

We will also use the following random variables.

$$\tilde{I}_1 = \{I_1^i \text{ with probability } q_i,\ i = 1, \ldots, n\}$$

$$\tilde{P}_1 = \{P_1^i \text{ with probability } q_i;\ i = 1, \ldots, n\}$$

The consumer allocates the available resources in each period between money and the consumption good. His behavior is described by a maximization of a utility function in which consumption and real balances appear as arguments:[3]

$$V(C_0, m_0, C_1^1, \ldots, C_1^n, m_1^1, \ldots, m_1^n).$$

---

[2] See Patinkin (1965) for a detailed framework of period analysis.

[3] The inclusion of real balances as an argument in the utility function is widely used since Samuelson (1947) and Patinkin (1965). This approach assumes that some important aspects of consumers' behavior can be captured without specifying the exact nature of the services which are provided by money.

It is assumed that the utility function is additive and that the consumer maximizes expected utility, thus

$$V(\ ) = U(C_0, m_0) + \rho \sum_{i=1}^{n} q_i U(C_1^i, m_1^i),$$

where ρ is a subjective rate of discount The maximization is done subject to a single budget constraint in period 0 and an anticipated budget constraint for period 1·

$$\max_{C_0, M_0, C_1{}^i, M_1{}^i \geq 0} U\left(C_0, \frac{M_0}{P_0}\right) + \rho \sum_{i=1}^{n} q_i U\left(C_1^i, \frac{M_1^i}{P_1^i}\right)$$

$$\text{s.t } (a)\ C_0 + \frac{M_0}{P_0} = W_0, \tag{1}$$

$$(b)\ C_1^i + \frac{M_1^i}{P_1^i} = W_1^i = \frac{I_1^i}{P_1^i} + \frac{M_0}{P_1^i}\ (i = 1, \quad , n)$$

Assume that $U(\ )$ is strictly quasi-concave to define

$$F(X) = \max_{C, M} U\left(C, \frac{M}{P}\right)$$

$$\text{s t. } C + \frac{M}{P} = X$$

$$\text{and } C, M \geq 0,$$

where $F(\cdot)$ is the maximum second period's utilities that the consumer can get from an amount of real resources X in the second period

There are two sources of benefits from holding real balances (a) "utility" in the present (which might stand for the "transaction motive" for holding money), and (b) uses in the future (the "store of value" function of money) When the consumer holds point anticipations (I_1, P_1), the value of holding M_0 today for future use, measured in period 1's utility units, is

$$F\left(\frac{M_0 + I_1}{P_1}\right) - F\left(\frac{I_1}{P_1}\right)$$

When future magnitudes are uncertain, the expected value of holding M_0 for future use is

$$E\left[F\left(\frac{M_0 + \tilde{I}_1}{\tilde{P}_1}\right) - F\left(\frac{\tilde{I}_1}{\tilde{P}_1}\right)\right]$$

The consumer weighs the total benefits from holding money (present plus future use) against the benefits of consumption today. Formally, this decision can be described by:

$$\max_{C_0, M_0 \geq 0} U\left(C_0, \frac{M_0}{P_0}\right) + \rho E\left[F\left(\frac{M_0 + \tilde{I}_1}{\tilde{P}_1}\right)\right] \qquad (2)$$

$$\text{s.t. } C_0 + \frac{M_0}{P_0} = W_0.$$

At a solution of (2) the expected marginal utility from holding one dollar as money (present and future uses) equals the marginal utility of a dollar's worth of present consumption. Thus,

$$U_2\left(\bar{C}_0, \frac{\bar{M}_0}{P_0}\right)\frac{1}{P_0} + \rho E\left[F'(\tilde{W}_1)\frac{1}{\tilde{P}_1}\right] = U_1\left(\bar{C}_0, \frac{\bar{M}_0}{P_0}\right)\frac{1}{P_0}. \qquad (3)$$

Where $(\bar{C}_0, \bar{M}_0)$ is a solution to (2), $U_i(\cdot\cdot)$ denotes partial derivatives and

$$\tilde{W}_1 = \frac{\bar{M}_0 + \tilde{I}_1}{\tilde{P}_1}.$$

Multiplying (3) by P_0 and substituting yields:

$$\begin{aligned} A(\bar{m}_0, W_0) &= U_1(W_0 - \bar{m}_0, \bar{m}_0) - U_2(W_0 - \bar{m}_0, \bar{m}_0) \\ &= \rho E\left(F'(\tilde{W}_1)\frac{P_0}{\tilde{P}_1}\right), \end{aligned} \qquad (4)$$

where $\bar{m}_0 = \bar{M}_0/P_0$ and the first equality defines the function $A(\ )$ as the difference between the marginal utility of consumption and the marginal utility of money in the first period.

In what follows we specialize in assuming that the consumer is risk neutral Thus $F'(\cdot)$ is a constant. Let $k = \rho F'(\cdot)$, where k is the "utility" of a unit of real wealth in the future, and define the real rate of return on money to be

$$\tilde{R} = \frac{M_0/\tilde{P}_1 - M_0/P_0}{M_0/P_0} = P_0/\tilde{P}_1 - 1.$$

Then (4) can be rewritten as

$$A(\bar{m}_0, W_0) = k[1 + (E\tilde{R})]. \qquad (5)$$

Reducing consumption in the present by one unit implies having an additional unit of real balance today and having $1 + \tilde{R}$ more units of real wealth in the future. Since (5) implies that the difference in the marginal utility between present consumption and present real balances is equal to the expected utility of $1 + \tilde{R}$ units of real wealth in the future, it insures that at the optimum the above change in allocation will not increase the expected utility.

The partial derivative of $A(\ \cdot)$ with respect to real balances is

$$\frac{\partial A(\ \cdot)}{\partial \bar{m}_0} = -U_{11}(\ \ ) + 2U_{12}(\ \ ) - U_{22}(\ \ ),$$

which is positive if we assume that the second order conditions of the problems are satisfied. Since $A(\ \ )$ is increasing in real balances, an increase in the expected real rate of return on money increases the demand for money

General equilibrium considerations which are developed in the second chapter of Eden (1975) leads to the above result without using the assumption of risk neutrality Specifically, it is shown that, even if all consumers have aversion to risk, the mean of the distribution of the real rate of return on money is the only moment of the distribution which effects the equilibrium quantity of money

III

Previous studies attempted to explain changes in the demand for money by using anticipation with respect to the expected rate of inflation rather than the expected real rate of return on money To appreciate the implications of this misspecification, we turn now to examine the relationship between the real rate of return and the rate of inflation

The discrete rate of inflation (for the period 0 -- t) [4]

$$\tilde{\pi} = \frac{\tilde{P}_t - P_0}{P_0}$$

is related to the real rate of return by

$$1 + \tilde{R} = \tilde{P}_0/\tilde{P}_t = \frac{1}{1 + \tilde{\pi}} \tag{6}$$

Since the function $1/(1 + \pi)$ is convex in π, Jensen's inequality combined with (6) implies that the expected real rate of return on money is greater when the rate of inflation is random compared to a nonrandom rate of inflation (holding the mean constant)

$$1 + E\tilde{R} = E\frac{1}{1 + \tilde{\pi}} \geq \frac{1}{1 + E\tilde{\pi}} \tag{7}$$

To illustrate inequality (7) consider the following example let $P_0 = 2$ and consider the following alternatives for the price level in period t

$$\text{Alternative 1}\quad P_t = 2 \text{ with probability } 1$$

$$\text{Alternative 2}\cdot\ \tilde{P}_t = \begin{Bmatrix} 3 \text{ with probability } 1/2 \\ 1 \text{ with probability } 1/2 \end{Bmatrix}.$$

---

[4] For a similar discussion on the continuous rate of inflation, see Eden (1975)

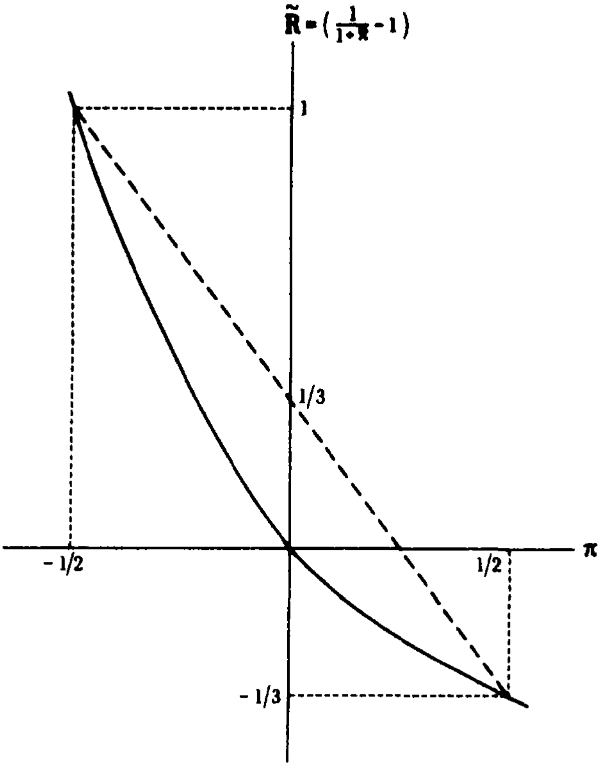


FIG 1 —The relationship between the expected real rate of return on money and the rate of inflation

Under both alternatives the expected rate of inflation equals zero. The expected real rate of return under alternative 1 is

$$P_0/P_t - 1 = 2/2 - 1 = 0$$

and under alternative 2

$$\begin{aligned} E(P_0/P_t - 1) &= 1/2(2/3 - 1) + 1/2(2/1 - 1) \\ &= -1/2\tfrac{1}{3} + 1/2\cdot 1 = \tfrac{1}{3}. \end{aligned}$$

Thus, the expected rate of return is higher for the random rate of inflation. This result is derived graphically in figure 1.

Further, it can be shown that the expected real rate of return increases with the variance of the rate of inflation (holding higher order moments constant). Therefore, if the demand for money is misspecified by using the expected rate of inflation rather than the expected real rate of return, the anticipated variance will show a significant positive effect [5]

Empirically, the distinction between the real rate of return and the rate of inflation is important only when there are changes in the anticipated variance of the rate of inflation It is shown in Eden (1975) that during the post-World War I periods of hyperinflations, changes in the variance of the rate of inflation were substantial and therefore the performance of the real rate of return in explaining changes in the demand for money was significantly better than the performance of the rate of inflation (in Cagan's [1956] original study)

References

Cagan, P. "The Monetary Dynamics of Hyperinflation " *Studies in the Quantity Theory of Money*, edited by M Friedman Chicago Univ Chicago Press, 1956

Eden, B. "Aspects of Uncertainty in Simple Monetary Models " Ph.D dissertation, Univ. Chicago, June 1975

Friedman, M. "The Quantity of Money—a Restatement " *Studies in the Quantity Theory of Money*, edited by M Friedman Chicago Univ Chicago Press, 1956

Patinkin, D *Money, Interest and Prices* New York Harper & Row, 1965

Samuelson, P. A. *The Foundations of Economic Analysis* Boston Harvard Univ. Press, 1947

Tobin, J. "Liquidity Preference as Behavior to vard Risk " *Rev Econ Studies* 25 (February 1958) 65–86

---

[5] It was argued that the distinction between the real rate of return and the rate of inflation will disappear for certain functional forms of the demand for money Since any monotone function of the expected rate of return will increase with the variance of the rate of inflation, and a function of the expected rate of inflation will not, this assertion is not correct

Vindication of a "Common Mistake" in Welfare Economics

John C. Panzar and Robert D. Willig

Bell Laboratories

In a recent note in this *Journal* (Dorfman 1975), Robert Dorfman mistakenly condemns a useful analytical device commonly employed in welfare economics. He contends that it is, in general, improper to derive the conditions necessary for Pareto efficiency from those necessary for the maximization of an arbitrary linear social welfare function

$$W = \sum_i w_i U^i, \qquad w_i \geq 0.$$

"Such a solution to the Pareto-efficiency problem will not necessarily satisfy the necessary conditions for a solution to the maximization problem, so those conditions are not necessary for Pareto efficiency" (p. 863). Professor Dorfman's criticism is unfounded, and those (unnamed) authors whom he reproves have made, at worst, a semantic error.

Let $x = (x_1, \ldots, x_m)$ be a vector characterizing social states, and let $U^i(x)$, $i = 1, 2, \ldots, n$, be differentiable utility functions representing the preferences of the agents over the states. Suppose there are constraints on the states given by the differentiable functions $F^t(x) \geq 0$, $t = 1, 2, \ldots, T$.[1] The work of S. Smale (1974*b*),[2] in a more general mathematical setting,[3] implies that a necessary condition for x^* to be Pareto optimal, subject to the constraints, is that there exist nonnegative numbers $w_1, \ldots, w_n, \lambda_1, \ldots, \lambda_T$, not all zero, such that

$$\sum_{i=1}^{n} w_i \nabla U^i(x^*) + \sum_{t \in S} \lambda_t \nabla F^t(x^*) = 0, \tag{1}$$

where $S = \{t \mid F^t(x^*) = 0\}$, and $F^t(x^*) \geq 0$ for $t = 1, \ldots, T$.

---

[1] These can be economic or technological constraints, including any needed nonnegativity constraints on the x_i's.

[2] We are grateful to an anonymous referee and to Carl Futia for bringing this work to our attention.

[3] Smale allows x to take values in an arbitrary manifold and characterizes the "first-order Pareto set" which includes the set of classical Pareto optima.

[*Journal of Political Economy*, 1976, vol. 84, no. 6]
© 1976 by The University of Chicago. All rights reserved.

The necessity of this condition for Pareto efficiency can be derived in a more restrictive context by familiar means. The Lagrangian for the Pareto program,

$$\max U^1(x) \text{ subject to } U^i(x) \geq U_0^i \quad i = 2, \ldots, n$$
$$F^t(x) \geq 0 \quad t = 1, \ldots, T, \tag{2}$$

is

$$L = U^1(x) + \sum_{i=2}^{n} w_i[U^i(x) - U_0^i] + \sum_{t=1}^{T} \lambda_t F^t(x).$$

The Kuhn-Tucker necessary conditions are (1), with $w_1 = 1$, and, additionally,

$$U^i(x^*) - U_0^i \geq 0, \; w_i[U^i(x^*) - U_0^i] = 0 \quad \text{for } i = 2, \ldots, n. \tag{3}$$

The condition (1) exactly characterizes the *critical*, or *stationary*, points of the linear social welfare function

$$W \equiv \sum_{i=1}^{n} w_i U^i(x)$$

over the feasible set. Thus, Smale's result implies that a necessary condition for x^* to be Pareto optimal is that there exist a nondegenerate linear social welfare function (with nonnegative weights) for which x^* is a stationary point.

Such points need not be maxima, but may, instead, be local minima or saddle points In figure 1, let the curve represent the utility possibility frontier for feasible choices of x in a two-person economy The values of x which generate points A and B are Pareto efficient in that they are true maxima of $U^1(x)$, subject to $U^2(x)$ being not less than certain numbers. However, while A corresponds to a local maximum of *some* linear social welfare function, B corresponds to a local minimum (This was the import of Dorfman's example)[4] Nonetheless, both A and B are stationary points of some linear combinations of the utility functions.

Dorfman has made a subtle error He is correct in writing: "Positively stated, it is quite possible for a set of x's to be Pareto efficient and still not maximize W for any positive w's" (p. 863). However, his conclusion (quoted earlier), that the first-order conditions for the maximization of some linear social welfare function are not necessary for Pareto efficiency, is false. These conditions, which are equivalent to (1), are, indeed, necessary, as is (3).

The authors rebuked by Dorfman may have mistakenly stated that they were maximizing W over some feasible set. Yet, their analyses have proceeded from correct first-order necessary conditions.

---

[4] Smale (1974*a*) also gives such an example.

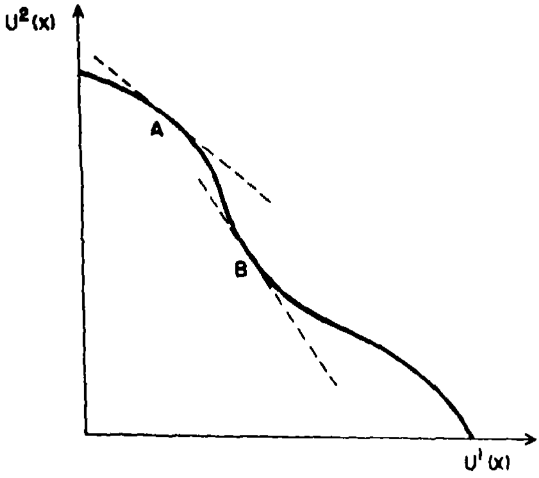


Fig 1

We conclude with the now obvious caution that if *second-order* conditions are to be invoked in a study of Pareto efficiency, they must be developed from the true program (2), and not from any other short-cut maximizations

References

Dorfman, Robert. "Note on a Common Mistake in Welfare Economics " *J P E* 83, no 4 (August 1975) 863–64

Smale, S "Global Analysis and Economics III Pareto Optima and Price Equilibria." *J Math Econ* 1 (August 1974) 107–18 (*a*)

——— "Global Analysis and Economics V Pareto Theory with Constraints " *J. Math. Econ.* 1 (December 1974) 213–21 (*b*)

Rejoinder

Robert Dorfman

Harvard University

In the note that stimulated Panzar and Willig's illuminating comment, I cautioned against an appealing but fallacious chain of reasoning that occurs frequently when writers are seeking conditions for Pareto-efficient resource allocations Formally, Pareto efficiency is an instance of vector maximization. But vector maximization is an awkward mathematical problem To avoid it, the writers in question replace the original Pareto-efficiency problem by the more ordinary problem of maximizing the synthetic objective function[1] $\sum w_i U^i(x)$ subject to the same constraints They reason, usually implicitly, that any Pareto-efficient allocation must be a solution to the synthetic problem for some set of weights, w_i This is simply incorrect (though it need not lead to incorrect conclusions) My example showed that, there were an infinite number of Pareto-efficient allocations but only two of them corresponded to possible maxima of the synthetic problem

Panzar and Willig and a number of private correspondents point out correctly that there is a saving feature The first-order conditions for x^* to be a local solution to the vector maximization problem are that the constraints be satisfied and that there be nonnegative numbers δ^* and nonnegative multipliers λ^* such that

$$\sum_{i=1}^{m} \delta_i^* \frac{\partial U^i(x^*)}{\partial x_j} + \sum_{i=1}^{T} \lambda_i^* \frac{\partial F^i(x^*)}{\partial x_j} = 0, \qquad j = 1, 2, \quad , m$$ [2]

These conditions are virtually the same as the analogous conditions for a solution to the synthetic problem But that does not justify the objectionable reasoning In my example we have the uncomfortable situation that all Pareto-efficient allocations can be discovered by *minimizing* the synthetic objective function, but that happenstance is surely an unappealing justification for introducing the synthetic problem

Panzar and Willig use a different, and more careful, device for reducing the vector maximization problem to an ordinary maximization problem,

---

[1] Panzar and Willig's notation is used throughout

[2] This theorem depends on the Kuhn-Tucker theorem and therefore presumes that the constraint qualification is satisfied at x^*

[*Journal of Political Economy*, 1976, vol 84, no 6]
© 1976 by The University of Chicago. All rights reserved.

but it too has its shortcomings. Besides, being asymmetrical, it is unsafe, because consumer number 1 may be special in some essential respects; he may be a vegetarian or even an ascetic. To see how this method can go astray you need only consider what would happen to it if consumer number 1 can be satiated with a smaller allocation than is left over after all the other consumers have attained their prescribed utility levels.

My recommendation is. If you have a vector maximization problem grit your teeth and face it for what it is

I want to thank Panzar and Willig, and my private correspondents, for clarifying my unduly cryptic note

Book Reviews

The Statistical Movement in Early Victorian Britain. The Foundations of Empirical Social Research. By MICHAEL J. CULLEN.
New York: Barnes & Noble, 1975. Pp. xii+205. $22.50.

The 1830s and the 1840s were marked by an unusual upsurge of interest in statistics in Great Britain. The Board of Trade's Statistical Department dates from 1832 and the General Register Office from 1836. The defense departments were commissioning a series of statistical reports on the health of the armed forces. Edwin Chadwick was using the resources of the Poor Law administration to conduct surveys of the health and sanitary conditions of the working classes. And in the early 1830s a host of statistical societies (or plans for statistical societies) sprang up in London and the provinces. Of these last, only the Statistical Society of London (much later the Royal Statistical Society) and the Manchester Statistical Society survived as effective institutions for the promotion and dissemination of statistical research. But in the 1830s there were also statistical societies in Belfast, Birmingham, Bristol, Glasgow, Leeds, Liverpool, and Newcastle actively conducting surveys of the condition of the working classes or of their educational standards.

This slim volume (originating from a Ph.D. dissertation and still devoting a quarter of its space to notes and references) documents and explains this sudden craze for statistics. It is not a critique of the statistics. Dr. Cullen is concerned less with the specific results of the movement—though, as he shows, these were substantial in total—than with the philosophy and motives of the leading personalities and with the way their personal commitments affected the range, bias, and quality of their achievement.

After a concise preliminary review of the work of previous British statisticians, from the political arithmeticians to John Rickman, the author describes and evaluates the role played by the activists in the statistical movement of the 1830s and 1840s, first in their official capacities and then as members of the local statistical societies. He finds G. R. Porter, for example, and William Farr, turning to the Statistical Society of London when their government departments cramped their statistical ambitions. On the other hand the Manchester Statistical Society lost much of its creative momentum when James Kay went to London to the Poor Law Commission and its other leading members diverted their reforming energies to the anti-corn-law league. For in spite of the prestigious support enjoyed by the early statistical societies their work (which turned out more expensive than their founders had dreamed) depended on the driving enthusiasm of a few active professionals.

Cullen accounts for the statistical movement as a response to the strains and stresses of the rapid urbanization taking place in the second quarter of the nineteenth century. Its leaders showed little interest in the broader social or economic consequences of industrialization. What really concerned them were its environmental implications, and they focused in particular on the high rates of crime and of mortality associated with urbanization. Above all, they were social reformers, determined to select and manipulate their statistics to justify the public health and educational prescriptions they regarded as the answer to the manifest deterioration in standards of health and morality in urban areas. Most of the energies and financial resources of the statistical societies were accordingly devoted to social surveys focused on the condition of the working classes and state of education. Within this frame of reference the questions they asked, and the answers they stressed, were conditioned by a strong moralistic middle-class ideology. "Indeed from many of the surveys of the statistical movement it is possible to deduce a better description of the books and pictures of the early Victorian working classes than of their income and expenditure" (p. 137).

The overriding concern of statisticians with social reform seems to have fallen away in the 1850s—partly no doubt because the prosperity of the mid-Victorian boom was reducing social tension, partly because the active statisticians were becoming more professional in their interests, and partly also because other political issues, less likely to be furthered by statistical propaganda, were coming to the fore The "statistical movement" as Cullen has interpreted it ends with the deaths of G. R. Porter and Joseph Fletcher in 1852.

This is a fascinating little book in spite of a rather terse style. The author's interpretation is cogent, if debatable, and he has managed to attach a considerable degree of coherence and breadth of interest to what may seem at first sight a limited topic.

Phyllis Deane

University of Cambridge

Domestic Monetary Management in Britain, 1919–1938 By Susan Howson University of Cambridge Department of Applied Economics, Occasional Paper 48. Cambridge: Cambridge University Press, 1975 Pp. 213

Any work that sheds new light on the economic history of interwar Britain is bound to be welcome, and Dr Howson's prize-winning dissertation is an important addition to the literature She has had access to treasury and cabinet documents of the period, only recently made available to research workers. She is able to give a much clearer picture than were earlier writers of the evolution of the opinions and aims of policymakers and how these related to the policies actually implemented. This strand in Howson's work is skillfully interwoven with an account of the effects of policy, notable, particularly in contemporary British literature, for the amount of emphasis that it lays on the role of monetary factor in influencing the level of economic activity.

Apart from its introduction and conclusion and a number of appendices, the book consists of five substantive chapters dealing, respectively, with the immediate postwar boom and slump, the period 1924–29 whose most notable event was the return to gold at prewar parity, the great depression of 1929–31 that culminated in the abandonment of gold, the cheap money era of 1932–35, and finally 1936–38, years marked both by a recession and the beginning of the rearmament program.

I found the account of the postwar boom and slump hard to follow, largely because Howson assumes rather too much knowledge on the part of her readers of the conduct of economic policy during the war. However, thereafter she tells her story clearly. In the wake of the violent fluctuations of the immediate postwar years monetary policy became geared to achieving and then maintaining an exchange rate that overvalued sterling Restoration and maintenance of the prewar parity was the overriding goal of the Bank of England The Treasury was intermittently concerned at the effects that the high short-term interest rates necessary to the achievement of this aim might have on unemployment, but did not question the appropriateness of the exchange rate target Indeed throughout the 1920s the Treasury exacerbated the unemployment problem by its attempts to lengthen the maturity of the national debt Hawtrey's doctrine that long-term interest rates were largely irrelevant to the behavior of aggregate demand was more an excuse for this policy than a cause of it, the Treasury viewed funding as an end in itself The great depression that began in the United States in 1929 ultimately forced Britain off gold in 1931 Treasury officials lamented this event almost as much as did those of the Bank, but they nevertheless used the new-found control over monetary policy that it gave them to institute a regime of cheap money consciously geared to promoting higher income and employment Howson argues convincingly that this policy had considerable success, especially in promoting the 1930s building boom Of course the Treasury did not acquire a new and coherent set of views on demand management policies overnight What Howson calls "the funding complex" persisted into the 1930s, and it was only later in the decade that the "treasury view" of fiscal policy began to be replaced within that institution by Keynesian ideas In the early 1930s monetary policy alone had to bear the burden of promoting an economic recovery that in the event was interrupted by the imported 1937 downturn By then rearmament was getting under way, and it was not until after the war that fully fledged Keynesian policies came into their own

Howson writes primarily as an economic historian and confines explicit economic analysis to her introductory chapter and an appendix on the term structure of interest rates Nevertheless the economics that underlies her work is coherent enough She puts more emphasis on the role of market interest rates in the transmission process of monetary policy than this reviewer would regard as appropriate She is well aware that there are no absolutes when it comes to judging what is or is not a "high" or "low" level for interest rates, but it would have been helpful had she made use of Fisher's distinction between real and nominal rates of return in dealing with these matters The inflation rate did after all vary quite a lot over the 2 decades with which she deals More discussion of the determinants of the banks' reserve ratios and the public's currency-deposit ratio would also have been helpful Fluctuations in these ratios played an important part in influencing the behavior of the money supply, particularly in the 1920s, and one would like to know much more than Howson tells her readers about why they behaved as they did. She is largely content to describe the sources of variations in the money supply when someone who lays more stress than does she on the latter variable would want these variations explained

All this is to say that Howson's monetary economics owes much more to Keynes than to Friedman or Fisher and not everyone would regard this as cause for criticism; but in any event the foregoing matters are relatively minor when set against what to this reviewer is the overriding virtue of her analysis namely, her understanding of the fundamental interdependence of the nature of the exchange rate regime and the conduct of domestic monetary policy in an open economy Howson never loses sight of this interconnection or permits her readers to do so One need only look at current debates about economic policy in Britain to realize what a scarce commodity is such understanding among British economists

This book, then, deals with work of high quality and it deserves a wide readership. It is therefore sad that so little thought seems to have gone into its manner of presentation. There are some minor blemishes: for example, a page full of yield curve diagrams (p. 171) whose axes are not even labeled let alone scaled, and references to otherwise unexplained events such as "the Hatry scandal," but these are merely irritating to the reader. Utterly infuriating is the lack of an index, and just as bad is the method used to identify the many quotations used in the text. References are given in footnotes at the back of the book, but these do not contain dates of publication; the latter are to be found in the bibliography. Who said what, and when he said it, is an integral part of Howson's story. With well over 100 footnotes per chapter, a bibliography that covers seven pages, and no index, it is excruciatingly difficult for the serious reader to follow this aspect of her narrative Also the lack of a uniform style for references in the bibliography (for journal articles page numbers are never given, and volume numbers seldom, for books the place of publication is sometimes missing and in one case the date as well) ensures that anyone wishing to obtain the material referred to here by way of an interlibrary loan service will need a lot of patience. I do not know how much blame for all this should go to Howson, and how much to the Cambridge University Press, but it is sincerely to be hoped that whoever is responsible will take note and resolve to do better next time. There are not so many good books written that we can afford to have their usefulness vitiated in this way by careless editorial work

DAVID LAIDLER

University of Western Ontario

Controls and Inflation The Economic Stabilization Program in Retrospect. By MARVIN H KOSTERS in association with J DAWSON AHALT (foreword by GEORGE P SHULTZ)
Washington American Enterprise Institute, 1975. Pp 135 $3 50 (paper)

It is hard to think of an economic policy decision in the twentieth century that aroused more excitement and consternation than the imposition of mandatory wage and price controls on August 15, 1971. Two years have passed since the controls program was terminated, and we are in a better position to assess the nature and significance of what really happened. This book by Marvin Kosters, who served as associate director for economic policy for the Cost of Living Council, provides us with the views of an insider. The book discusses controls policies in relation to the overall economic magnitudes (prices, wages, and profits) and also examines their implementation and impact in specific sectors (food, lumber, petroleum, cattle hides, and fertilizer). While some attention is given to the effects of controls on economic efficiency and the relationship of controls to demand management policies, the book is more chronological and institutional than analytical, and no sharp conclusions seem to stand out. Indeed, in many ways the things not explained in the book are more interesting than those that are.

One thing that is noticeably absent is a statement providing a rationale for controls as an economic policy tool. To be sure there are hints of such a rationale sprinkled here and there, but these generally amount to suggestions that controls dampen inflationary expectations or help to moderate excessive negotiated wage settlements. Such claims cannot be accepted as obviously true, however. The assertion that controls affect inflationary expectations ignores the theory of rational expectations which has played an important role in economic analysis in recent years and hence needs to be defended on both theoretical and empirical grounds In a similar vein it is not unreasonable to ask for some fuller explanation of just how inflation is related to what happens at the bargaining table

This book conveys the strong impression that the 1971–74 controls policy was not, in fact, based on any well-thought-out theory of inflation [1] In particular, the arguments presented certainly do not appear to be related to a theory of inflation based on the supply and demand for money Very little is said about the relationship of the controls and the policies of the Fed, other than to suggest that (retrospectively) the money supply grew too rapidly in 1971 and 1972 And even this is explained as a failure of economic forecasting rather than as an error induced by the controls [2] Even less is said about the effect of controls on the demand for money, although this would appear to be a matter of utmost importance in any policy intended to deal with inflation. This is a dramatic contrast to the writings of such well-known monetarists as J K Galbraith, who, in his book *A Theory of Price Control*, focusses substantial attention, at least implicitly, on the relationship of controls and the demand for money (1952)

The book's failure to rationalize the controls as a tool of economic policy may be explained on the grounds that they were initiated in response to political rather than economic conditions Kosters does refer to political pressures at several points but completely ignores the important question of just how the politically influential stood to gain from the controls policy He occasionally seems to imply that the political pressure for controls arose out of ignorance on the part of the public as to what the controls would do Ignorance is never a very satisfactory explanation for social behavior, and it surely does not explain why the administration requested an extension of mandatory authority beyond April 30, 1974, when congressional and public support for such an extension was clearly minimal

In view of the abysmal performance that most observers now ascribe to the controls, it is puzzling that Kosters is not more decisive in his condemnation of them Before the controls were imposed, one of the arguments the Nixon administration used against them was that they would not really contain inflation and that a "catch-up" burst of price rises was likely when they were relaxed. Instead of noting the essential validity of that prediction, Kosters seems to go out of his way to argue that the ineffectiveness of the Phase III controls in the first half of 1973 should not be viewed as a failure of the controls policy Moreover, despite the widespread evidence of economic dislocations and the heated complaints about inequities, Kosters blandly asserts that the short-run costs of controls were apparently not enormous He follows this remark with the claim that the costs of controls are nonetheless real but difficult to quantify (p 100) The implication is that costs which are difficult to quantify are apparently not enormous This conclusion should be great comfort to those involved in cost-benefit analyses

One of the great concerns that many economists have about incomes policies is the extent to which they involve the government in decision making that is best left to the private sector. We find evidence in this book that such involvement was more extensive than many may have realized Kosters notes "Adequacy of rates of return and willingness to make new investment commitments were factors considered in decontrol decisions. Securing capacity expansion commitments as controls were removed was part of an intricate process to facilitate orderly sectoral decontrol. Investment commitments provided a supporting rationale for sequential control decisions, and they represented a significant effort to coordinate policies for achieving capacity expansion needs with policies for removing controls" (p. 106). Thus the reward of decontrol (or sanction of continued control) was used to motivate capacity expansion and investment in specific industries and firms above the amount that presumably would have taken place voluntarily. Kosters does not explain how this policy relates to inflation or how the controllers knew the private sector's capital needs better than the business community. There is also a serious issue as to whether such policy is an appropriate use of the power vested in the controllers—consider the above paragraph with the words "political contributions" substituted for the phrases "investment commitments" and "capacity expansion commitments."

---

[1] This impression is supported by Arnold Weber's remarks in his review of the 1971 wage-price freeze: "The need for surprise should not preclude some planning for the implementation of a wage-price freeze It is true that the administration's shift in economic policy was initiated with an urgency that did not permit extensive planning of policies and procedures The basic decisions were made by the top officials of government —including the President—in a 48-hour period The total government commitment to planning an economic stabilization program at the time of the freeze was one middle-level civil servant in OEP" (1973, p 124)

[2] See the remarkably brief section on controls and demand management (pp 102–3)

Those who believe that wage and price controls are a costly and unjustified government activity will find nothing in this book to shake their conviction. The same may be said for those who believe the contrary.

John P. Gould

University of Chicago

References

Galbraith, J K *A Theory of Price Control.* Cambridge, Mass.: Harvard Univ Press, 1952.

Weber, A. R. *In Pursuit of Price Stability The Wage Price Freeze of 1971.* Washington Brookings Inst., 1973.

The Joyless Economy. An Inquiry into Human Satisfaction and Dissatisfaction. By Tibor Scitovsky.
New York· Oxford University Press, 1976. Pp. xviii+310. $11.95.

In this book, Professor Scitovsky has combined two related monographs Each could have been published separately, but each draws strength from the other. The first is entitled "The Psychology and Economics of Motivation." The line of argument is that economists have developed models of human behavior which explain, at most, a minor fraction of human goal-seeking activity. He contends that physiological psychology has more to contribute than economics toward an understanding of the phenomena with which economists ought to be concerned; and the bulk of the first half of the book contains a résumé of his understanding of physiological psychology in its relation to the concerns of the economist. He states the point of view taken in the first part:

> Economists . accept unquestioningly the consumer's judgment of what is best for him, his tastes as the outcome of that judgment, and his market behavior as the reflection of his tastes. Economists will not analyze the motivation of consumer behavior, claiming that that would be beyond their competence and that it would inevitably involve an improper judgment: that of judging one man's way of making the best of his life by another man's standards. As a result, the economist's approach to his

> subject tacitly assumes that consumers know what they are doing and are doing the best they can, so that the economist's only task is to see to it that the economy delivers what consumers want
>
> That approach overlooks the fact that tastes are highly variable, easily influenced by example, custom, and suggestion, constantly changed by the accumulation of experience, and modified by changing prices and availability of some satisfactions and the unavailability of others It also overlooks the possibility that the same influences that modify our tastes might also modify our ability to derive satisfaction from the things that cater to our tastes In short, the economist's standard procedure of postulating that each consumer knows best what is good for him and trusting the consumer's behavior to reflect that knowledge seems to be unscientific. What I propose to do instead is to look into the consumer's behavior, analyze its motivation, and then draw whatever conclusions follow from such an investigation [Pp 3–4]

He does not prove his case. To do so, were the case correct, would require many lifetimes of work and far more printed pages than are in the book under review What he does do very successfully is to raise issues which are fundamental and not easily answered

Part of what is said in both halves of the book draws upon distinctions between "pain," "comfort," and "novelty," and how they are related to the enjoyment of life. Comfort should be thought of as the relief of pain or discomfort Comfort is enough to eat, a soft chair, etc. The way economists model utility, with tastes unchanging and more always assumed to be better, suggests that more comfort is always an improvement Scitovsky argues that in North America, Europe, and Japan (the relatively rich countries) nearly everyone has about all the comfort he (does or should?) want. For awhile, the achieving of ever-higher levels of comfort can be much fun Then the novelty wears off, and the comforts achieved are taken for granted. The consumer gets hardly more pleasure out of his 60-foot yacht than he got years before out of his 18-foot sloop Of course, this suggests that one's past pattern of consumption affects one's current pleasure Or, put another way, tastes do change. This is hardly news, for any economist, indeed any fellow human, must know this The questions of interest are Does the current state of consumer theory capture some valuable insights? Can the economist's understanding and modeling of the consumer be improved with the help of the work Scitovsky cites?

The issue is not settled by the glib assertion that current satisfaction may be written as a function of previous levels of consumption, that is, that in the tth period of the consumer's life $u_t = u_t(c_1, \ldots, c_t)$ Consideration of the sources and meaning of novelty should make this clear As the word implies, novelty resides in what is new and unexpected. A complex piece of music is novel for many hearings because new aspects may be noticed with each performance Several live performances of the same piece present some novelty because they are not identical, etc. Novelty, according to the author, is the real stuff of human satisfaction How do we get that into the utility function?

The second part of the book is called "The American Way of Life." This part is much more fun to read and makes better cocktail-party conversation, however, to me it seems rather less successful The heart of the argument here is that we Americans purchase more comfort and less novelty than we should Were we to shift our consumption patterns we would be happier.

There are aspects of the living style of certain Europeans which appeal to many American academics. Indeed, both American and European academics tend toward agreement, which suggests that those with a disposition to become professors also have a disposition to like such things as art works, old buildings and towns, varied food, and so forth. Though the tastes approved by the author in this part of the book cater to my prejudices too, I fear that its whole thrust is that everyone would be better off if he changed his consumption habits to those of Tibor Scitovsky. The case is argued by him in a gracious and undogmatic way. It also is the more convincing due to the preparation from the first half of the book Anyway, perhaps he is right.

Before closing, I want to add some summary comments of the author's·

> Even at my most ambitious, I could only hope to open up a new field of enquiry, raise new questions, and provide a little reassurance that those questions are not as intractable as they first seem This book is open ended, partly because the subject is open ended. If I have provoked some readers into continuing, extending, testing, contradicting my argument, I have done well.
>
> The novelty of this book lies in introducing novelty as an object of desire and a source of satisfaction. To devote a whole book to the trivial task of adding one more source of satisfaction to the economist's already long list would be absurd, if it were not for the fact that the stimulus of novelty is among the most fundamental of human needs, and that novelty is a rather special commodity, as special as virginity or a delicate flower, and very unlike the economist's stock in trade. My fellow economists have been too earnest and puritanical to recognize the consumer's need for novelty, and the American consumer himself is also too earnest and puritanical to admit to himself his own desire for stimulation and novelty. We could say that there is nothing wrong with seeking comfort rather than stimulus if that is one's ideal of the good life, except for the following reasons. First, we overindulge in comfort, second, in the modern economy, the economies of scale impose the majority's tastes on the whole society, and when the majority chooses to sacrifice the stimulus of novelty for the sake of comfort, the creation of novelty and the minority's seeking new ways of attaining the good life are both impeded. [Pp. 282–83]

Whether this book could, should, or will have any impact on the development of economics remains to be seen, however, it is a thoughtful, well-written, seriously intended piece of work which does raise questions which are difficult, important, and interesting. I am pleased to have read it, and I recommend it.

JAMES W. FRIEDMAN

University of Rochester

Did Monetary Forces Cause the Great Depression? By PETER TEMIN. New York: W. W. Norton & Co., 1976. Pp. 201. $8.95.

That feud is warming up again: monetarists and antimonetarists, Chicago and Cambridge. If Temin had wanted to keep the peace, the title could have been *Consumers May Have Started the Great Depression.*

Temin's preface states as his purpose "the application of modern macroeconomic tools" to analysis of the disaster in 1930–33 (p xii) [1] He intends to discriminate precisely between two classes of explanation, the "money hypothesis" and the "spending hypothesis," that emerged from earlier probes into the unprecedented decline of economic welfare Allegedly, advocates of neither view have constructed and tested their models precisely, and they have talked past one another, failing in resolution of their dispute Temin set out to construct explicit models, to test them, and to initiate eyeball-to-eyeball confrontations between monetarists and their critics. He finishes the book on a note of disappointment and frustration "The movements of the observed variables in the interwar period are too limited to provide enough information to test alternative macroeconomic hypotheses about them" (p 178) The limited tests that can be used lead him to the conclusion that the spending hypothesis is "more plausible" than its alternative The note of frustration is loudest in the final sentence "The economist who uses this conclusion or any other conclusion about the Depression as a basis for economic policy recommendations essentially is performing an act of faith" (p 178)

Temin's spending hypothesis marks 1930 as the year of commitment to deep depression A small recession started in 1929, "due to some combination of factors which cannot be disentangled" (p 170) The stock-market crash had something to do with it, along with excess stocks of housing, crop failures, and monetary restraint. Recovery might have come quickly if two "unexplained events" had not intervened in 1930 and if European financial crises in 1931 had not provoked financial stringency and shocked spenders' expectations The major "event" was an autonomous fall in consumption amounting to $3 billion, about 3 percent of GNP. The decline is autonomous in the sense that it is the residual in estimates of consumption by life-cycle functions relating nominal consumer expenditures during 1919–41 to nominal disposable income and nominal net household wealth The residual for 1930 exceeds by a wide margin residuals for other recession years of the period. An alternative estimate, of real consumption regressed on real permanent income, also yields a substantial negative residual for 1930 The minor "event" was a decline, attributable to developments abroad, of approximately $1 billion in American exports Before the banking crisis of 1930, the economy was committed to a depression by these autonomous declines of $4 billion in spending

Europe's crisis in 1931, beginning with failure of Austria's Credit Anstalt, and the Federal Reserve's response with restrictive monetary policies accelerated the momentum of depression Europe was in trouble partly because of domestic distortions, partly as a result of recession in the United States It is likely that any feedback upon this country would have been slight if the Federal Reserve had not chosen in Autumn 1931 to protect its gold position by raising interest rates and tightening financial markets. After 1931, the depression raced relentlessly to its climax in 1933 during political stalemate in this country and under pressures that were worldwide and that cannot be analyzed from narrow national perspectives. Nothing, not endogenous factors such as deflation of money-wage rates and intermittent growth in real money balances, not exogenous factors such as monetary and fiscal policies, emerged before the New Deal to stop recession of output and deflation of prices

Temin's spending hypothesis is less precise about the economy's processes of recession and deflation than about the autonomous causes. He lists a number of them, but he refrains from measurement because there is no reliable macroeconometric model for the American economy in 1919–41 Such a model is beyond reach because of data limitations and because economic processes were distorted by random shocks. Models based upon data for recent decades are not relevant, because the economy's structure has changed. They are frail frameworks for analysis even of economic change during 1945–76 since these years "have not encompassed enough variation for alternate theories to be tested" (p. 177). Macroeconometrics in the style of large models may be fun and games, but it cannot yet probe objectively the processes of growth and instability.

---

[1] Unless otherwise specified, all page references are to the book under review

"Unexplained events" were the impulse that started the Depression according to both the spending hypothesis of Temin and the money hypothesis of Friedman-Schwartz (Friedman and Schwartz 1963, pp. 299–419). The Temin events were the declines in consumption and exports in 1930. The Friedman-Schwartz (F-S) event was a banking and liquidity crisis in late 1930 (pp. 10, 15). In both hypotheses, small rocks start landslides, but the rocks are different (Friedman and Schwartz 1963, p. 419).

Temin hammers away at the money hypothesis and F-S. Three indicators persuade him that monetary restraint did not cause the depression. "The monetary base (high-powered money) continued to increase during the Depression except for a brief dip in 1930" (p. 9). "There is no evidence in the interest rates of the monetary stringency cited in the money hypothesis" (p. 126). "The real stock of money did not fall in these years. If the demand for money can be written in real terms . . . , then there does not seem to have been any contractionary pressure on real income from the supply of money" (p. 166).

The money hypothesis stipulates excess demand for money after the bank failures of late 1930 and an equilibrating process that involved first, tightness in financial markets and then falling employment, output, and price levels. In Temin's judgment, the F-S narrative assumes excess money demand, but "it does not test it or prove it at all" (p. 16). The spending hypothesis suggests, and Temin demonstrates to his satisfaction, that there was excess supply of money except when the Federal Reserve intervened restrictively before the stock-market crash in 1929 and during the European crisis in 1931. In the one view, the supply curve in the market for money receded along the demand curve, inducing depression and deflation. In the other view, it was the demand curve that receded under pressure of declining real income and prices.

Temin objects to F-S analytical techniques including the familiar expression $M = H(D/R[1 + (D/C)])/[D/R + D/C]$ that relates the money supply (M) to high-powered money (H), banks' deposit-reserve ratio (D/R), and the public's deposit-currency ratio (D/C) (pp. 16–17, 27). It defines the stock of money. It is not a supply equation, and it is not joined with a demand equation to determine equilibrium supply in the market for money. When one specifies a behavioral model of the market for money, it becomes apparent that (D/R) and (D/C) are complementary means by which falling demand for money reduces the stock of money. The role of banking failures in the Great Depression, in Temin's judgment, was to reduce these ratios and to drain excess supply from the market for money (pp. 9–10, 20).

Temin dismisses the counterfactual "thought experiments" that F-S engage in. (Friedman and Schwartz, pp. 303–5, 313, 343; Temin, p. 28). He condemns the casual way of F-S with history. At one point, F-S attribute bank failures in late 1930 to economic distress in agricultural areas, at another to poor loans and investments in the 1920s (pp. 85–90). The F-S explanation for failure of the Bank of United States in 1930 is denied: fraud was the reason for it, not a crisis of liquidity. The F-S "view of the Bank of United States is completely at variance with the facts" (p. 91).

Friedman-Schwartz and such architects of macromodels as Klein and Tinbergen are not the only targets in Temin's shooting gallery Others have proposed explanations of the Great Depression that rely on unobserved and unpredictable magnitudes and cannot be tested (pp 174–75) Accounts that stress declines in investment in this country or in flows of capital to Germany have the facts wrong (pp. 152, 172). The notion that World War I touched off a long swing or Kuznets cycle, with the Great Depression as its final phase, has not been substantiated, and the lag structure that would link the political to the economic disaster probably cannot be specified (p. 52) No one has a cause of the depression that can compete with that residual of $3 billion in Temin's estimate of the consumption function.

Temin's little book is a fascinating exercise in the style of modern economic history. Dissent from some parts of it here is not intended to turn readers away One basis for dissent is that Temin finds excess supplies on all markets, except for the commercial-paper market at times, during the depression Outputs, labor, physical wealth, foreign exchange, banks, nominal money, swamped their respective markets Did everyone offer to sell some of everything in exchange for nothing? Is the notion of a budget constraint irrelevant for 1930–33? Temin observes that, "to the extent that they [people] bought other assets, like U.S government securities or diamonds, they were reducing their total cash balances—as opposed to simply changing their composition We do not know what people bought when they 'sold' bank deposits" (p 54) Excess supplies prevailed on all markets ex ante and floated away into some wide blue yonder ex post.

Temin explains price-level deflation in 1930–33 by "the same arguments used to explain the fall in income" (p 168) The deflation was not propelled by money except in late 1929 and in late 1931 It was exogenous to the market for money. Deflation did feed back into that market, reducing demand for nominal money and, by such processes as bank failures, reducing supply of nominal money as well. Temin's indicator for money's passive role is the interest rate for prime commercial paper That is the price that clears the market for money, in his judgment, and its downward drift proves that money was in ex ante excess supply This analysis will have monetarists pounding walls with their fists They will grant that deposits with weak banks were in excess supply They will insist that there was excess demand for deposits with safe banks, currency postal savings deposits—and, among other things, prime commercial paper Prime commercial paper, moreover, was in demand by banks for rediscount, since the Federal Reserve was not providing reserves by other means The commercial-paper rate is not the simple test for disequilibrium on the market for money that Temin assumes it to be.

Temin suggests another feedback from deflation to the market for money deflation maintained the stock of real balances Since the stock was stable during 1929–33, he contends, there could hardly have been excess demand for it (p. 166). Now a stable stock is not firm evidence of an equilibrium stock and of continuous tranquillity on the market for money. Tossing aside the commercial-paper rate as indicator of excess supply in that market, the monetarist is bound to insist that constant real balances were achieved only by chronic and severe deflationary pressure from excess money demand As Temin concedes, the money hypothesis is still breathing, but it does need a little more oxygen

E S SHAW

Stanford, California

Reference

Friedman, Milton, and Schwartz, Anna Jacobson *A Monetary History of the United States 1867–1960*. Princeton, N J Princeton Univ. Press, 1963

A Theory of Profits. By Adrian Wood.
Cambridge: Cambridge University Press, 1975. Pp. 184. $14.50 (cloth).

The major thrust of *A Theory of Profits* is to develop an analytical set of tools that will enable a student of the modern corporation to understand the firm's financial activities. Wood believes that this is necessary because existing theory, based on such assumptions as perfect markets, wealth-maximizing management, interest rates compensating lenders for the riskiness of borrowers, etc., has certain limitations. The assumptions of the dominant neoclassical view are, in his opinion, "wildly misleading. They convey an image of the capital market . . . which is unrealistic to the point of almost complete irrelevance" (p. 6).

It is Wood's belief that finance matters. A firm's capital structure is not trivial (because individuals can arrange their own leverage). A firm's dividend payout decision is not a matter of indifference (because if earnings are retained, stock prices will rise and the individual can sell as many shares as he wishes to determine his own dividend income). Stock prices do not always sell at book value. A firm does not adopt all investment projects whose expected returns are greater than the cost of capital and then determine how to raise the funds for the acceptable projects. Rather, there is a continuous interplay between the availability of funds from internally generated sources and investment projects.

Wood's basic model is developed in the first three chapters of this five-chapter book This analysis rests on the interplay of four ratios as the firm tries to maximize its sales The basic ratios are · (1) the gross retention ratio (r), the ratio of retained earnings and depreciation charges to gross operating profits (i e , gross revenues less variable operating costs, profits [P] are therefore defined as the firm's internally generated cash flow before interest payments, dividend payments, and taxes), (2) the external finance ratio (X), the ratio of new borrowing and new equity issues to gross plant and equipment plus inventory, (3) the financial asset ratio (f), the ratio of liquid assets to total investment, and (4) the investment coefficient, k, the ratio of new investment outlays (I) to the change in sales ($V - \bar{V}$).

A firm's total cash outlays will be $(1 + f)I$ Part of this flow can be financed externally, XI, the remainder must be financed internally. By definition,

$$rP = (1 + f - X)I$$

and

$$P = \frac{(1 + f - X)I}{r} . \tag{1}$$

Since

$$I = kgv, \tag{2}$$

where sales growth is defined as $(V - \bar{V})/V$ and a firm's profit margin is defined as

$$\pi = P/V, \tag{3}$$

equation (1) can be rewritten as

$$\pi = \frac{(1 + f - X)}{r} gk. \tag{1'}$$

Much of chapter 2 is spent in describing how the values of f, X, and r are fixed by institutional forces and why, once set, they are difficult to change. The impact of equations (1) and (1′) then is that the level of investment depends directly on the profits generated by a firm and the growth rate on the firm's profit margin.

Placed in an operational context, equation (1′), which is called the "finance frontier" can be written as

$$\pi \geq \frac{(1 + f - X)}{r} gk \qquad (1'')$$

Thus, at all times, the margin must be adequate to finance expansion

Firms also face an investment opportunity schedule,

$$\pi \leq u(g, k), \qquad (4)$$

where

$$\partial\pi/\partial g < 0 \qquad (4a)$$

and

$$\partial\pi/\partial k > 0 \qquad (4b)$$

Condition (4a) is the "demand-profit margin tradeoff" It reflects the fact that, once on an opportunity frontier, the firm must "raise its unit selling costs and keep down its prices in order to increase the level of demand for its products in the face of competition from other firms" (p 70)

Condition (4b) "reflects the fact that once the opportunity frontier has been attained, increasing the level of investment expenditure has a beneficial effect upon the demand-profit margin tradeoff as a result, for example, of reductions in unit operating costs" (p 70)

The problem of a firm, then, is to maximize g subject to two constraints, (1'') and (4) The three requirements between them determine the firm's profit margin, π, the rate of growth of sales, g and the investment coefficient, k

Wood's framework of analysis, then, is straightforward and constructive It integrates a number of traditional financial decision variables in a rational and coherent way The power of the framework, however, depends on the assumptions that f, X, and r are fixed or fluctuate within narrow ranges While much of chapter 2, as noted above, is spent describing the kinds of considerations that would suggest that these ratios are constant, the question is an empirical one and no data are presented in this work

Chapter 4 recasts the investment function of macroeconomic models in light of microfinancial policies and, for me, held no surprises Chapter 5 concludes the book with some implications for government policy and did have some conclusions I did not expect

Basing my opinion on the model Wood developed in the first three chapters, I would have expected that the policy recommendations would be consistent with increasing a corporation's internal cash flow, that is, profits as Wood defines the term. After all, this is how investment, growth, greater employment, more leisure, etc., will come about Instead, I find a concern with the implications of the model for the distribution of income The problem that bothers Wood is that "the ownership of capital, since it involves very little work, is an inadequate moral basis for receiving income which would otherwise accrue to wage and salary earners [Moreover] the ownership of company securities is so highly concentrated that a reduction in the share of profits would make the distribution of income among persons significantly less unequal and would therefore cause income to be distributed more nearly in accordance with needs" (p 155)

Wood is therefore torn between two positions On the one hand, he is sympathetic to the strength of the capitalistic desire for growth He quotes from Marx "Accumulate! Accumulate! That is Moses and the prophets" (p 12) On the other hand, he "accepts without question that other things being equal it is desirable to reduce the shares of dividends, interest and retained earnings in national income" (p 158). His resolution of the dilemma he poses for himself is the classical academic plea "To understand the way in which the economy works is a legitimate end in itself" (p 155)

EUGENE LERNER

Northwestern University

Index to Volume 84

Articles and Notes

Abrams, Burton A, and Settle, Russell F. The Effect of Broadcasting on Political Campaign Spending An Empirical Investigation 1095

Atkinson, Scott E, and Halvorsen, Robert Interfuel Substitution in Steam Electric Power Generation 959

Barro, Robert J Reply to Feldstein and Buchanan 343

Barzel, Yoram. An Alternative Approach to the Analysis of Taxation 1177

Baumol, William J It Takes Two to Tango, or *Sind* "Separable Externalities" *Überhaupt Möglich?* 381

Becker, Gary S, and Tomes, Nigel Child Endowments and the Quantity and Quality of Children S143

Ben-Porath, Yoram Fertility Response to Child Mortality Micro Data from Israel S163

Bergstrom, T C Toward a Deeper Economics of Sleeping 411

Bitros, George C A Statistical Theory of Expenditures in Capital Maintenance and Repair 917

Black, Stanley W. Multilateral and Bilateral Measures of Effective Exchange Rates in a World Model of Traded Goods 615

Blackford, George H Money, Interest, and Prices in Market Disequilibrium. A Comment 893

Blinder, Alan S, and Weiss, Yoram Human Capital and Labor Supply: A Synthesis 449

Brenner, Menachem A Note on Risk, Return, and Equilibrium: Empirical Tests 407

Bronfenbrenner, Martin Economics in Dialectical Dialect 123

Brown, Charles A Model of Optimal Human-Capital Accumulation and the Wages of Young High School Graduates 299

Browning, Edgar K The Marginal Cost of Public Funds 283

Bruce, Neil See *Harberger, Arnold C*

Buchanan, James M Barro on the Ricardian Equivalence Theorem 337

Burgess, David F Tariffs and Income Distribution Some Empirical Evidence for the United States 17

Cain, Glen G, and Dooley, Martin D Estimation of a Model of Labor Supply, Fertility, and Wages of Married Women S179

Campbell, J. M., Jr. See *Gregory, Paul R*

Cheh, John H A Note on Tariffs, Nontariff Barriers, and Labor Protection in United States Manufacturing Industries 389

Chen, Chau-nan. Currency Denominations and the Price Level 179
Chiswick, Barry R. The Effect of Unemployment Compensation on a Seasonal Industry: Agriculture 591
Christensen, Laurits R., and Greene, William H. Economies of Scale in U.S. Electric Power Generation 655
Clark, Colin. Economic Development in Communist China 239
Connolly, Michael, and Taylor, Dean. Testing the Monetary Approach to Devaluation in Developing Countries 849
Cox, Charles C. Futures Trading and Market Information 1215
Darby, Michael R. Three-and-a-Half Million U.S. Employees Have Been Mislaid: Or, an Explanation of Unemployment, 1934–1941 . . 1
Das, S. K. See *Murphy, J. C.*
Della Valle, Philip A., and Oguchi, Noriyoshi. Distribution, the Aggregate Consumption Function, and the Level of Economic Development: Some Cross-Country Results . 1325
De Vany, Arthur. Uncertainty, Waiting Time, and Capacity Utilization: A Stochastic Theory of Product Quality 523
Deyak, Timothy A., and Smith, V Kerry The Economic Value of Statute Reform. The Case of Liberalized Abortion 83
Dooley, Martin D. See *Cain, Glen G.*
Dorfman, Robert. Rejoinder . . . 1365
Dornbusch, Rudiger. Expectations and Exchange Rate Dynamics 1161
Douglas, Paul H. The Cobb-Douglas Production Function Once Again: Its History, Its Testing, and Some New Empirical Values 903
Durden, Garey C. See *Silberman, Jonathan I*
Eden, Benjamin. On the Specification of the Demand for Money. The Real Rate of Return versus the Rate of Inflation 1353
Evenson, Robert E., and Kislev, Yoav A Stochastic Model of Applied Research 265
Fama, Eugene F. Inflation Uncertainty and Expected Returns on Treasury Bills . . . 427
Feige, Edgar L , and Pearce, Douglas K. Economically Rational Expectations· Are Innovations in the Rate of Inflation Independent of Innovations in Measures of Monetary and Fiscal Policy? . . 499
Feldman, Roger. Some More Problems with Income-contingent Loans. The Case of Medical Education . 1305
Feldstein, Martin. Perceived Wealth in Bonds and Social Security: A Comment . . . 331
Feldstein, Martin. Temporary Layoffs in the Theory of Unemployment 937
Findlay, Ronald, and Wellisz, Stanislaw. Project Evaluation, Shadow Prices, and Trade Policy . . 543
Frech, H. E., III. The Property Rights Theory of the Firm Empirical Results from a Natural Experiment 143
Fry, Maxwell J. A Purchasing-Power-Parity Application to Demand for Money in Afghanistan 1133
Galai, Dan Policy Implication of Full-Loss-Offset Capital Gains Tax on Forward Contracts. A Note on Leland's "Optimal Forward Exchange Positions" 1313
Garbade, Kenneth D., and Silber, William L. Price Dispersion in the Government Securities Market 721
Gordon, Scott. The New Contractarians 573
Green, Jerry, and Sheshinski, Eytan. Direct versus Indirect Remedies for Externalities 797
Greene, William H. See *Christensen, Laurits R.*

Gregory, Paul R., and Campbell, J. M , Jr Fertility Interactions and Modernization Turning Points . . . 835
Griliches, Zvi Wages of Very Young Men . S69
Gronau, Reuben. The Allocation of Time of Israeli Women S201
Gruen, Arthur. An Inquiry into the Economics of Race-Track Gambling 169
Halvorsen, Robert See *Atkinson, Scott E*
Hamada, Koichi. A Strategic Analysis of Monetary Interdependence 677
Hamilton, Bruce W The Effects of Property Taxes and Local Public Spending on Property Values A Theoretical Comment 647
Hansen, W Lee See *Knapp, Charles B.*
Hanson, John R , II. The Leff Conjecture Some Contrary Evidence . 401
Harberger, Arnold C , and Bruce, Neil The Incidence and Efficiency Effects of Taxes on Income from Capital· A Reply 1285
Heckman, James J A Life-Cycle Model of Earnings, Learning, and Consumption S11
Hondai, Susumu Market Organization and the Effect of Import Quotas Empirical Evidence 1123
James, John A. The Conundrum of the Low Issue of National Bank Notes 359
Johnson, Harry G Destabilizing Speculation A General Equilibrium Approach 101
Jones, Robert A The Origin and Development of Media of Exchange 757
Jonson, P D Money and Economic Activity in the Open Economy. The United Kingdom, 1880–1970 979
Kapur, Basant K Alternative Stabilization Policies for Less-developed Economies 777
Kihlstrom, Richard E. Firm Demand for Information about Price and Technology . 1335
Kislev, Yoav See *Evenson, Robert E*
Klevmarken, Anders, and Quigley, John M Age, Experience, Earnings, and Investments in Human Capital 47
Knapp, Charles B , and Hansen, W Lee Earnings and Individual Variations in Postschool Human Investment 351
Latham, Joe. Triangular Graphs 623
Lindsay, Cotton M A Theory of Government Enterprise 1061
McKenzie, George W Measuring Gains and Losses 641
Marvel, Howard P The Economics of Information and Retail Gasoline Price Behavior An Empirical Analysis 1033
Meadows, George Richard Taxes, Spending, and Property Values A Comment and Futher Results 869
Mincer, Jacob Unemployment Effects of Minimum Wages S87
Mitchell, Bridger M , and Phelps, Charles E National Health Insurance Some Costs and Effects of Mandated Employee Coverage 553
Munch, Patricia An Economic Analysis of Eminent Domain 473
Munnell, Alicia H Private Pensions and Saving New Evidence 1013
Murphy, J. C , and Das, S K Money Illusion and Balance-of-Payments Adjustment 73
Oguchi, Noriyoshi See *Della Valle, Philip A*
Oi, Walter Residential Location and Labor Supply S221
Panzar, John C , and Willig, Robert D Vindication of a "Common Mistake" in Welfare Economics 1361
Pashigian, B. Peter Consequences and Causes of Public Ownership of Urban Transit Facilities . 1239
Pearce, Douglas K See *Feige, Edgar L*

Phelps, Charles E. See *Mitchell, Bridger M.*

Pryor, Frederic L. The Friedman-Savage Utility Function in Cross-cultural Perspective 821

Quandt, Richard E. Some Quantitative Aspects of the Economics Journal Literature 741

Quigley, John M. See *Klevmarken, Anders.*

Rees, Albert. H. Gregg Lewis and the Development of Analytical Labor Economics S3

Rosen, Sherwin. A Theory of Life Earnings S45

Rosenberg, Nathan. Another Advantage of the Division of Labor 861

Salant, Stephen W. Exhaustible Resources and Industrial Structure: A Nash-Cournot Approach to the World Oil Market 1079

Sampson, Gary, and Yeats, Alexander. Do Import Levies Matter? The Case of Sweden 881

Sargent, Thomas J. A Classical Macroeconometric Model for the United States 207

Sargent, Thomas J. The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics 631

Schwartz, Aba. Migration, Age, and Education . 701

Settle, Russell F. See *Abrams, Burton A.*

Shavell, Steven. Sharing Risks of Deferred Payment . . . 161

Sheshinski, Eytan. See *Green, Jerry.*

Shoven, John B. The Incidence and Efficiency Effects of Taxes on Income from Capital . . . 1261

Silber, William L See *Garbade, Kenneth D.*

Silberman, Jonathan I, and Durden, Garey C. Determining Legislative Preferences on the Minimum Wage An Economic Approach 317

Skogh, Göran. The Social Costs of Monopoly and Regulation: Some Comments . . . 1319

Smith, V Kerry See *Deyak, Timothy A.*

Stern, Robert M. Some Evidence on the Factor Content of West Germany's Foreign Trade . . 131

Stigler, George J. The Successes and Failures of Professor Smith . 1199

Strickland, Allyn D., and Weiss, Leonard W. Advertising, Concentration, and Price-Cost Margins 1109

Swan, Peter L. Optimum Replacement of Capital Goods with Labor-saving Technical Progress. A Comparison of the Early New England and British Textile Firm 1293

Tarascio, Vincent J. Pareto: A View of the Present through the Past . . 109

Taylor, Dean. See *Connolly, Michael.*

Tideman, T. Nicolaus, and Tullock, Gordon. A New and Superior Process for Making Social Choices 1145

Tomes, Nigel. See *Becker, Gary S.*

Toyoda, Toshihisa. Import Instability and Tariffs: Some Welfare Implications of Price Stabilization 395

Tullock, Gordon. See *Tideman, T. Nicolaus.*

Vining, Daniel R., Jr. Autocorrelated Growth Rates and the Pareto Law: A Further Analysis. 369

Vermani, Arvind. A Dynamic Model of the Firm. 603

Weiss, Laurence. The Desirability of Cheating Incentives and Randomness in the Optimal Income Tax 1343

Weiss, Leonard W. See *Strickland, Allyn D.*

Weiss, Yoram. See *Blinder, Alan S.*

Welch, Finis. Employment Quotas for Minorities S105
Wellisz, Stanislaw. See *Findlay, Ronald.*
Whalley, John. Thornton's Estimates of Efficiency Losses in Soviet Industry
  Some Fixed-Point-Method Recalculations 153
Williamson, Jeffrey G Technology, Growth, and History 809
Willig, Robert D. See *Panzar, John C*
Yeats, Alexander. See *Sampson, Gary*

Reviews

Allingham, Michael. *Equilibrium and Disequilibrium A Quantitative Analysis of Economic Interaction* By R M Witcomb 413
Cullen, Michael J *The Statistical Movement in Early Victorian Britain The Foundations of Empirical Social Research* By Phyllis Deane 1367
Douglas, George W, and Miller, James C, III *Economic Regulation of Domestic Air Transport Theory and Policy.* By William A. Jordan 185
Fair, Ray C. *A Model of Macroeconomic Activity* Vol. 1 *The Theoretical Model* By Uri M Possen 189
Helleiner, Karl F *Free Trade and Frustration Anglo-Austrian Negotiations, 1860-1870.* By Victoria Curzon 895
Howson, Susan *Domestic Monetary Management in Britain, 1919-1938.* By David Laidler 1368
Knudsen, Odin, and Parnes, Andrew *Trade Instability and Economic Development An Empirical Study* By R M Sundrum 897
Kosters, Marvin H, in association with J Dawson Ahalt *Controls and Inflation · The Economic Stabilization Program in Retrospect* By John P Gould 1370
Labys, Walter C, ed. *Quantitative Models of Commodity Markets* By Charles C Cox 651
McKenzie, Richard B, and Staaf, Robert J *An Economic Theory of Learning* By Edward Lazear 192
Manne, Henry G, and Miller, Roger Leroy, eds. *Gold, Money and the Law.* By Elmus R Wicker 1139
Munro, John H A *Wool, Cloth, and Gold · The Struggle for Bullion in Anglo-Burgundian Trade, 1340-1478* By Frank Spooner 194
Nadiri, M Ishaq, and Rosen, Sherwin *A Disequilibrium Model of Demand for Factors of Production* By Frank Brechling 652
Parkin, Michael, and Sumner, Michael T, eds *Incomes Policy and Inflation.* By Edgar L Feige 415
Pasinetti, Luigi L *Growth and Income Distribution Essays in Economic Theory* By Duncan K. Foley 195
Pedersen, Jørgen. *Essays in Monetary Theory and Related Subjects* By Jürg Niehans 1141
Phillips, Almarin, ed. *Promoting Competition in Regulated Markets* By Michael M. Murphy 417
Plant, Arnold. *Selected Economic Essays and Addresses* By Almarin Phillips 1142
Prachowny, Martin F. J. *Small Open Economies Their Structure and Policy Environment.* By Harry G. Johnson 420
Sandberg, Lars G. *Lancashire in Decline A Study in Entrepreneurship, Technology, and International Trade.* By Donald N. McCloskey 198

Scitovsky, Tibor. *The Joyless Economy: An Inquiry into Human Satisfaction and Human Dissatisfaction.* By JAMES W. FRIEDMAN 1372
Sowell, Th[illegible]. *Classical Economics Reconsidered.* By SAMUEL HOLLANDER [illegible]
Spence, [illegible]ichael. *Market Signaling: Informational Transfer in Hiring and Related S[illegible]ing Processes.* By EMMETT KEELER 200
Steele, E. [illegible]. *Irish Land and British Politics: Tenant-Right and Nationality, 1865-1[illegible].* By BARBARA L. SOLOW 420
Temin, [illegible]er. *Did Monetary Forces Cause the Great Depression?* By E. S. SHAW 1374
Tinbergen, Jan. *Income Distribution: Analysis and Policies.* By BARRY R. CHISWICK 422
Williamson, Jeffrey G. *Late Nineteenth Century American Development: A General Equilibrium History.* By GAVIN WRIGHT 201
Wood, Adrian. *A Theory of Profits.* By EUGENE LERNER 1378